2014

The design of adaptive clinical trials

Sasiwimon Iwsakul
University of Wollongong

Recommended Citation

Iwsakul, Sasiwimon, The design of adaptive clinical trials, Doctor of Philosophy thesis, School of Mathematics and Applied Statistics, University of Wollongong, 2014. https://ro.uow.edu.au/theses/4568

The Design of Adaptive Clinical Trials

A thesis submitted in fulfillment of the requirements
for the award of the degree

Doctor of Philosophy

from

University of Wollongong

by

Sasiwimon Iwsakul
M.Sc (Statistics), Chulalongkorn University

School of Mathematics and Applied Statistics

2014

Certification

I, Sasiwimon Iwsakul, declare that this thesis, submitted in fulfillment of the requirement for the award of Doctor of Philosophy, in the School of Mathematics and Applied Statistics, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Sasiwimon Iwsakul

August 30, 2015

Abstract

Clinical trials are research studies which involve both healthy people and patients. The aim of these trials is to assess the efficacy of a new treatment or to compare the efficacy of new treatments with the current treatment. In this thesis, we focus on comparing the efficacy of a new treatment with that of an existing treatment.

The objective of a clinical trial design is to obtain a correct conclusion as well as to be concerned with the economic issues. Since this trial involves human beings, importantly, it should address ethical concerns. In a traditional (equal randomisation, ER) design, half of the participants will be assigned to an inferior treatment. When comparing the efficacy of treatments, one of the disadvantages of the ER design is that although the evidence may be strong that one treatment is superior to others, a trial cannot finish early. As a result of the issues discussed above, ethical problems and economic issues arise. In order to cope with these problems, an adaptive design should be considered. An adaptive design is one in which a trial can be changed (adapted) during its progress. These changes are based upon the accrued data. In this thesis, the designs developed involve two specific areas of adaptive designs: adaptive randomisation and interim analyses. Particular attention is paid to response-adaptive and covariate-adaptive randomisations.

In this thesis, the designs of Huang *et al.* (2009) (HNL) are extended and generalised. The first step is to examine two aspects of these designs: (1) the enrolment regime, and (2) the randomisation procedure. We modify the recruitment regime from the adaptive method of HNL, by changing the arrival rate from exactly one patient per week to an average of one per week. In real life, it is rare to find that patients come into the trial at a rate of exactly one patient per week. Then, simulation is carried out to investigate how this more realistic scenario affects the results. It is found that the differences between the statistical properties of the two enrolment regimes (i.e. exactly one new patient per week, and an average of one new patient per week) should not be considered practically significant. We conclude that the HNL practice is a sensible approach to use, and follow their practice of having exactly one arrival per week.

We also investigate several important criteria for evaluating and comparing designs. We focus on the Operating Characteristic curve, and various design characteristics. It is found that the OC curve is not an appropriate method of comparing clinical trial designs due to the complication that the OC curve is not uniquely defined by the difference in treatment means.

In this thesis, eight design characteristics are considered. By using these design characteristics, in a simulation, we find that as far as economical and ethical reasons are concerned, the adaptive design that uses the response-adaptive (RA) randomisation is better than the ER design. If the main concern is statistical power, the RA design is a competitive design.

The HNL design is also extended to a much more applicable design. We intend to enable it to perform in a more realistic situation. We then develop an appropriate randomisation procedure by considering the response of the previous patients and the degree of covariate imbalance. Covariates are some prognostic factors which may be important when considering a cancer trial. Failure to take account of these covariates might lead to bias or the wrong conclusion. In order to decrease bias and provide more effective comparisons, in the proposed response-adaptive, covariate-adjusted (RACA) design, a subsequent patient will be more likely to receive the better treatment and the degree of covariate imbalance will be minimized. Covariates are incorporated into a trial by using a better procedure. That is, the various covariates are considered simultaneously and can be dependent upon one another.

Then the design characteristics and the degree of covariate imbalance are employed to compare the performances of the RACA design with those of the RA design. In a simulation, it is found that, as far as these characteristics are concerned, the performances of the RACA and RA designs are only slightly different. However, if the degree of covariate imbalance is of principal concern, the RACA design is superior to the RA design.

We conclude that the RACA design is the best design, and has the advantages for economic as well as ethical issues. It also gives a reasonable statistical power which is competitive.

At the conclusion of this thesis, we suggest further lines of research.

Acknowledgements

This thesis would not have been possible without the help and support of many people. First of all, I would like to sincerely thank, Professor Ken Russell for his support, guidance, encouragement and patience throughout this thesis. I would also like to thank my co-supervisor, Associate Professor Marijka Batterham for her support, additional suggestion and being a facilitator between my supervisor and I of a long distance communication.

In particular, I am very grateful for his and her belief in me that I would finish my thesis, when I had trouble believing myself. Additionally, thank both of you, Associate Professor Adam Rennie, Dr Ngamta Thamwattana and Rhondalee Cambalee for Sessional International tuition fee award.

I would like to acknowledge Professor David Steel and Dr Carole Birrell for allowing me to sit in their class. I have gained valuable knowledge and experience from the course which considerably contributed to the development of this work.

My deep appreciation to office staff of the School of Mathematics and Applied Statistics: Anica Damcevski, Ann Harper, Carolyn Silveri, Kerrie Gamble and Lisa Pyle for your help and generosity. Our daily conversations not only improved my English but also developed my confidence. I am very grateful for your attempts to help me overcome many obstacles and settle in Australia.

I would also like to thank to all my friends, particularly Dr Diane Hindmarsh and Dr Sarah Neville for our discussion and giving me good advice. Also thanks to Janet Russell and people in ICIS for their help. In particular, Bob Colvin who

has proofread some parts of my thesis.

Last but not least, I wish to express my gratitude to my family: mother, brothers, sisters, nieces and nephews, for all their love, encouragement, always believing in me and supporting a lot of money for me to study here.

Contents

1 Introduction 1

1.1 Clinical trials . . . 1

1.2 The design of clinical trials . . . 3

1.3 Adaptive design . . . 3

1.4 Bayesian Method . . . 5

1.5 The literature review . . . 6

1.6 Huang *et al.* (2009) . . . 15

1.6.1 Theory . . . 16

1.7 Overview of this thesis . . . 21

2 Investigation of the Huang *et al.* (2009) method 23

2.1 Huang *et al.* (2009) program . . . 23

2.1.1 Aspects of Ning (2009) . . . 24

2.1.2 Errors in Ning (2009) . . . 27

2.1.3 The results from Ning (2009) . . . 28

2.2 The recruitment regime . . . 34

2.3 Comparison of the HNL designs under two arrival rates . . . 37

2.3.1 Hypothesis Testing . . . 38

2.3.2 The Results of the Hypothesis Testing . . . 40

3 Evaluation of Clinical Trial Designs 43

3.1 Operating Characteristic curve . . . 43

3.2 Example of using the Operating Characteristic to compare designs 44

3.3 Design characteristics 47
3.3.1 Assumptions 50
3.4 Expected cost 50
3.5 Example of using design characteristics to evaluate designs 54
3.5.1 The results of the expected cost 62
3.6 Example of using design characteristics to compare clinical trial designs 66
3.6.1 The results of expected cost 73
3.7 The expected total costs of the designs 75
3.8 Example of using the expected total cost 81
3.8.1 Discussion 87
4 Extension of the HNL design 88
4.1 Introduction 88
4.2 Ning and Huang (2010) 89
4.2.1 The RACA design 90
4.2.2 Simulation 95
4.3 The Ning2 98
4.3.1 Differences 98
4.3.2 Results 99
4.4 Extension of the HNL design 110
4.4.1 RA randomization 111
4.4.2 CA randomization 112
4.4.3 RACA randomization 119
4.4.4 Simulation 119
4.5 Results 126
4.5.1 Hypothesis testing 136
4.5.2 A comparison of expected costs 146

4.5.3 A comparison of expected total costs 147

4.6 Conclusion 150

5 Conclusions 151

5.1 Summary 151

5.2 Discussion 155

5.3 Future research 158

A Main R program used 161

References 171

List of Figures

3.1 E(total cost) of the ER and HNL designs as a function of P_{H_0}. 85

4.1 The histograms of the number of patients assigned to arms A and B for the RACA design with $N_{max} = 120$. 131
  (a) the number of patients assigned to arm A 131
  (b) the number of patients assigned to arm B 131

4.2 The histograms of the number of patients assigned to arms A and B for the CA design with $N_{max} = 120$. 132
  (a) the number of patients assigned to arm A 132
  (b) the number of patients assigned to arm B 132

4.3 The histograms of the number of patients assigned to arms A and B for the RACA design with N_{max} of 100. 133
  (a) the number of patients assigned to arm A 133
  (b) the number of patients assigned to arm B 133

4.4 E(total cost) of the RACA and RA designs as a function of P_{H_0}. . . . 149

List of Tables

2.1 Comparison of the results for the HNL design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.975$, $N_{Max} = 120$ and $n_0 = 1$ 29

2.2 Comparison of the results for the HNL design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.975$, $N_{Max} = 120$ and $n_0 = 30$ 30

2.3 Comparison of the results for the common design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.993$, $N_{Max} = 120$ and $n_0 = 1$ 31

2.4 Comparison of the results for the common design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.993$, $N_{Max} = 120$ and $n_0 = 30$ 32

2.5 Comparison of the results of the HNL approach with a recruitment rate of exactly one patient per week with the results of the modified approach with the accrual rate being an average of one patient per week. 42

3.1 Comparison of the probabilities of retaining H_0 when H_0 is true for $N_{max} = 120$ under various scenarios 46

3.2 Comparison of the probabilities of Type I error for the HNL design, under various values of N_{max} 56

3.3 Comparison of the design characteristics when H_0 is false under various values of N_{max} when the treatment effect is 24.725 57

(a) Section 1 57

(b) Section 2 57

3.4 Comparison of the design characteristics when H_0 is false under various values of N_{max} when the treatment effect is 29.1 58

(a) Section 1 58

(b) Section 2 58

3.5 Comparison of the design characteristics when H_0 is false under various values of N_{max} when the treatment effect is 34.65 59

(a) Section 1 59

(b) Section 2 59

3.6 Comparison of the design characteristics of the HNL design when H_0 is false under various values of N_{max} when the treatment effect is 39.9 60

(a) Section 1 60

(b) Section 2 60

3.7 Comparison of the design characteristics of the HNL design when H_0 is false under various values of N_{max} when the treatment effect is 70.35 61

(a) Section 1 61

(b) Section 2 61

3.8 Comparison of some features of the HNL design, the common design for $n_0 = 1$ and the ER design 67

3.9 Comparison of the probabilities of Type I error for the HNL design, the ER and the common design when N_{max} is 120 71

3.10 Comparison of the design characteristics when H_0 is false for the HNL design, the ER design and the common design when the treatment effect is 70.35 and N_{max} is 120 72

(a) Section 1 72

(b) Section 2 72

3.11 The associated costs of the trial for four outcomes as given by using the definition of Parnell (2002). 78

3.12 The associated cost of the trial for the six outcomes as given in this research. 79

3.13 The associated costs of the trial for four outcomes. 81

4.1 The contingency table for Z_j that shows where the expected value in (4.3) comes from. 93

4.2 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 1 (where the efficacies of arms A and B are equal), when using $p_U = 0.95$, $N_{max} = 60$, and $n_0 = 10$. 101

4.3 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 1 (where the efficacies of arms A and B are equal), when using $p_U = 0.975$, $N_{max} = 100$, and $n_0 = 20$. 102

4.4 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with seed 1234, in Scenario 1 (where the efficacies of arms A and B are equal), when using different values of p_U, $N_{max} = 60$, and $n_0 = 10$. The values of p_U have been selected to give a probability of Type I error of approximately 0.10. 104

4.5 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with seed 1234, in Scenario 1 (where the efficacies of arms A and B are equal), when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05. 105

4.6 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 2 (arm B is better than arm A), when using different values of p_U, $N_{max} = 60$, and $n_0 = 10$. 106

4.7 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 2 (arm B is better than arm A), when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. 107

4.8 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2), when using different values of p_U, $N_{max} = 60$, and $n_0 = 10$. 108

4.9 Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2), when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. 109

4.10 An example of how to obtain this statistic in the case of a male smoker who is assigned to arm A 114

4.11 An example of assigning a patient by using the Frane method when there are two covariates 114

4.12 The contingency table for (Z_1, Z_2) 116

4.13 An example of the numbers of patients in each combination of levels of covariates in the two arms when there are triple covariates 116

4.14 The conditional probabilities for Z_2 121

4.15 The conditional probabilities for Z_3 122

4.16 The values of β_{0k}, β_{Tk}, β_{1k}, β_{2k} and β_{3k} 123

4.17 The values of $\mu_{x,k}$ in Scenarios 1 and 2 for all possible covariate combinations when k = 1, 2, 3, 4 when the parameters in Table 4.16 are used. 124

4.18 The values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when k = 1, 2, 3, 4, in Scenario 1 when the parameters in Table 4.16 are used. 125

4.19 The values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when k = 1, 2, 3, 4, in Scenario 2 when the parameters in Table 4.16 are used. 125

4.20 The values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when k = 1, 2, 3, 4, in Scenario 3 when the parameters in Table 4.16 are used. 127

4.21 Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 (where the efficacies of arms A and B are equal) when using $p_U = 0.95$, $N_{max} = 100$, and $n_0 = 20$. . . . 127

4.22 Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 (where the efficacies of arms A and B are equal) when using $p_U = 0.95$, $N_{max} = 120$, and $n_0 = 30$. . . . 128

4.23 Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 (where the efficacies of arms A and B are equal) when using $p_U = 0.95$, $N_{max} = 130$, and $n_0 = 30$. . . . 128

4.24 Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05. 129

4.25 Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 when using different values of p_U, $N_{max} = 120$, and $n_0 = 30$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05. 129

4.26 Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 when using different values of p_U, $N_{max} = 130$, and $n_0 = 30$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05. 130

4.27 Comparison of the results for the ER, RA, RACA and CA designs of Table 4.24, in Scenario 2 (arm B is better than arm A) when $N_{max} = 100$, and $n_0 = 20$. 140

4.28 Comparison of the results for the ER, RA, RACA and CA designs of Table 4.25, in Scenario 2 (arm B is better than arm A) when $N_{max} = 120$, and $n_0 = 30$. 141

4.29 Comparison of the results for the ER, RA, RACA and CA designs of Table 4.26, in Scenario 2 (arm B is better than arm A) when $N_{max} = 130$, and $n_0 = 30$. 142

4.30 Comparison of the results for the ER, RA, RACA and CA designs of Table 4.24, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2) when $N_{max} =$ 100,and $n_0 = 20$. 143

4.31 Comparison of the results for the ER, RA, RACA and CA designs of Table 4.25, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2) when $N_{max} = 120$, $n_0 = 30$. 144

4.32 Comparison of the results for the ER, RA, RACA and CA designs of Table 4.26, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2) when $N_{max} = 130$, $n_0 = 30$. 145

List of Abbreviations

AND	Average number of deaths
ANP	Average number of patients
ALT	Average length of the trial
ANPS	Average number of patients who gain a successful outcome from the treatments
CA	Covariate-adaptive
CR	Complete remission
ER	Equal randomization
HNL	Huang, Ning, Li, Estey, Issa, and Berry (2010)
NH	Ning and Huang (2010)
Ning2	A second Ning program
N_{Max}	Maximum sample size
n_0	Initial number of patients before adaptive randomization commenced
OC	Operating Characteristic
PBA	Percentage of patients assigned to the better treatment
PET	Probability of early termination
PR	Partial remission
RA	Response adaptive
RACA	Response-adaptive, covariate-adjusted
SDNPS	Standard deviation of the number of patients who gain successful outcome from the treatments

Chapter 1

Introduction

1.1 Clinical trials

Nowadays, clinical trials have widespread acceptance as a reliable method for assessing the efficacy and safety of a new treatment.

Clinical trials are research studies whose aim is to evaluate the efficacy of a new treatment or a new treatment procedure (e.g. new surgical procedure), including a new medical device (e.g. blood pressure gauge). These trials involve both healthy people and patients. For healthy people, clinical trials can be conducted to assess the safety of a new treatment. Moreover, for patients, these can be carried out to evaluate the advantages of a treatment. Clinical trials may also use to compare a new treatment against a treatment that is already available (Wang and Bakhai, 2006). Although there are several different kinds of clinical trials, the research in this thesis concentrates on Treatment Trials. These are clinical trials which test the efficacy of new treatments or new combinations of existing treatments.

Most clinical research is conducted in an orderly series of stages, called phases. There are four phases of clinical trials:

- Phase I trials

  This is the first phase, which involves human volunteers. An experimental treatment will be tested to decide the safe dose that can be given without toxicity and side effects. This phase requires a small group of people (20-80) and will last from several weeks to a few months (Pocock, 1993).

- Phase II trials

  The aim of Phase II trials is to investigate the efficacy and to further evaluate the safety of a treatment in a larger group of people (100-200). These trials will take longer time than Phase I trials. They might take several months to several years (Pocock, 1993).

  According to Yin (2012), this phase might be carried out to evaluate the efficacy of a treatment (single arm) or to compare the efficacy of several treatments (multiple arms). For a single arm, an experimental treatment will be evaluated and compared with the data from previous studies. When conducting multiple arms, some patients will be assigned to an experimental treatment whereas some will be assigned to a placebo or a standard treatment. Hence, conducting multiple arms might reduce a problem that occurs because of different situations between the present and previous studies, such as different populations and different doctors.

- Phase III trials

  If a new treatment is demonstrated to be effective in Phase II trials,, this treatment will proceed to Phase III trials. In these trials, a new treatment will be evaluated for its efficacy and safety in comparison to the current standard treatment. These trials recruit hundreds, even thousands, of patients and may continue from several months to several years (Pocock, 1993).

- Phase IV trials

  This phase is conducted after a treatment has been approved and released onto the market. The aim of this phase is to evaluate the long-term side effects of the treatment.

1.2 The design of clinical trials

An experimental design is a process of planning a study to obtain reliable data as well as a correct conclusion in order to answer the research questions effectively (Montgomery, 2001). In the design of an experiment, the main concern is to examine the effect of some intervention on the experimental units under controlled conditions. The design of an experiment has a very broad application in many disciplines, including clinical trials.

The design of clinical trials is the process of planning research to assess proposed treatments with respect to safety and efficacy. Although both Phase II and Phase III trials are involved with clinical trial design, this research focuses on Phase II trials. This is because Phase III studies require more resources, such as participants, time and finance. Phase II trials are much smaller and more manageable than Phase III trials.

The role of a statistician in the design of clinical trials covers many matters, including: determining the numbers of volunteers likely to be needed to detect a meaningful difference between the new product and the current treatment; devising procedures that can be used to decide whether a trial can be finished ahead of schedule if a difference between the new and current products becomes apparent very early in the trial; devising rules that determine which procedure the next volunteer should be allocated to; and evaluating and comparing a new design with competing designs to decide whether the new design is effective and competitive.

Of the four topics above, we are most interested in working on the last two.

1.3 Adaptive design

This section introduces an adaptive design, which is the main design to be discussed throughout this thesis. We will address some aspects of this design about which we will be concerned.

In recent years, adaptive designs have played a crucial role in clinical research and development. According to Dragalin (2006) and Chow *et al.* (2005), an adaptive design can be defined as a clinical trial design in which a trial can be modified by using data collected during the progress of the trial. This data is used to determine how to change a trial without affecting its validity and integrity. Adaptive designs are most useful in Phase II and Phase III trials.

The designs developed in this thesis involve two specific areas of adaptive designs:

1. Adaptive Randomisation

   By using this randomisation, the probability that a new patient receives a treatment varies depending on the collected data, instead of this probability being fixed over the period of the trial and equal randomisation being used (Dragalin, 2006). This randomisation is useful for clinical trials because, mostly, patients are recruited sequentially. In this research, we focus on the following two forms of randomisations:

   - Response-Adaptive Randomisation (or Outcome-Adaptive Randomisation)

     Response-Adaptive Randomisation is a randomisation in which the probability of assigning a patient to a treatment is based on the response of the previous patients. A higher proportion of patients are allocated to a better treatment (Korn and Freidlin, 2011). Consequently, this randomisation can provide ethical advantages over equal randomisation by decreasing the number of patients assigned to an inferior treatment.

   - Covariate Adaptive Randomisation

     According to **?**, covariate adaptive randomisation is a randomisation which is based on important prognostic factors that might have an influence on the response. These factors will be identified at the beginning of the trial. In this randomisation, a new patient will be assigned to a treatment by considering the balance of these factors so far.

2. Interim Analyses

   An Interim Analysis is a statistical analysis conducted during the progress of the trial to ensure both efficacy and safety. Interim Analyses allow for the possibility of early stopping. The trial will stop early if there is strong evidence that one treatment is superior or the new treatment is not worthwhile (futility) (Cook and DeMets, 2008).

For the reasons described above, using an adaptive design provides advantages for economical and ethical reasons. The economical reason is that by using interim analyses, an adaptive design can reduce the time required to show efficacy or futility. Decreasing the length of clinical trials results in lower drug development costs and reduced time to market. The ethical reason is that, by decreasing the length of trials, an adaptive design can enhance patient wellbeing. Moreover, since a smaller number of patients will be assigned to an inferior treatment, it will reduce the risk of receiving an inferior treatment.

1.4 Bayesian Method

In this research, we will produce a prior distribution $\pi(\theta)$ for each parameter θ, and update it using the information obtained from samples (y) to calculate the posterior distribution $p(\theta/y)$ of the parameter. The updating from the prior distribution to the posterior distribution is carried out using Bayes' theorem:

$$p(\theta/y) \propto f(y/\theta).\pi(\theta) = L(\theta; y)\pi(\theta), \tag{1.1}$$

where $f(y/\theta)$ is the sampling distribution of the response variable, $L(\theta; y)$ is the likelihood function and $\pi(\theta)$ is the prior distribution of θ (Mukhopadhyay, 2000, pp. 477 - 479).

1.5 The literature review

This section reviews the relevant literature on Bayesian Methods, survival analysis and adaptive designs.

In Huang *et al.* (2009), the authors suggested a new type of design for randomised clinical trials that includes the following advantages:

- Response-Adaptive Randomisation,
- Interim Analyses.

Another benefit of this design is that both survival and short-term patient response are used as primary endpoints. This is different from the approaches described in the earlier literature. Previously, researchers just used survival or short-term patient response as the primary endpoint. However, they did not use these endpoints together.

Although survival is the goal of clinical trials, this endpoint does not meet the requirement of response-adaptive randomisation. This randomisation requires the results of the previous patients immediately before assigning a treatment to the next patient. Since using short-term patient response can meet this requirement, it should be used as a primary endpoint as well.

The authors found that this new type of design can reduce the use of resources such as the number of patients and the time required for the trial. Additionally, more patients could be assigned to the better treatment.

The advantages of the Huang *et al.* (2009) design make it interesting, and it should be considered further. It will be described in more detail in the following section.

Berry (2004) supported the use of the Bayesian approach in adaptive designs. In these designs, modifications are made to a clinical trial during its progress by using collected data. Hence, the effectiveness of the design is based on the data obtained, and updated data will be required. The Bayesian method is a tool to obtain effective updated information, so it should be used in adaptive designs. In addition, Berry (2004) stated that the use of adaptive designs is increasing in cancer trials, not only in trials sponsored by pharmaceutical companies but also in many trials at The University of Texas M. D. Anderson Cancer Center (MDACC).

In Berry (2004), Bayesian decision-theory was used to examine the results from each possible sample size and identified the sample size that has the maximum expected utility or the minimum expected loss. Utility can be defined in terms of the efficacy of treatment for patients. In Huang *et al.* (2009), for example, utility is the mean progression-free survival time. We can also determine utility in terms of the economy, e.g., cost, time. A loss may occur from the wrong decision taken at the conclusion; for example, if treatment A is really better than treatment B but in a trial we decide that treatment B is superior to treatment A. The loss happens because patients will take treatment B instead of treatment A.

Berry also suggested that decision-theory can answer the ethical problems in clinical research. For ethical reason, patients should be effectively treated, whether they are patients who will receive benefits from the results during the trial or after the trials as a results of information gained from the trials.

For the reasons described above, the Bayesian method will be used in our thesis. Moreover, we aim to consider utility of designs in terms of both the efficacy of treatment and economy.

Let us consider the required sample size for a clinical trial. Dupont and

Plummer (1990) stated that factors that should be considered when selecting the sample size are the power of the test, economy, and time. Also, the researcher should balance the demand of these factors. Dupont and Plummer (1990) considered some tests such as log rank tests of survival data and t-tests for independent continuous response data. Peto *et al.* (1977) and Bewick *et al.* (2004) explained that the log-rank test is a procedure that is used to compare survival curves. We interpret this to mean that it can compare the efficacy of treatments in a clinical trial by looking at the survival times of patients under each treatment. Hence, these patients are observed for some specified period, or until death intervenes.

In addition, Clarke and Yuan (2006) considered the determination of sample sizes by the Bayesian method. According to Inoue *et al.* (2005), the aim of Bayesian sample size determination is to find the appropriate sample size to achieve the required goals. For instance, in decision-theory, researchers require the sample size that has the maximum expected utility or the minimum expected loss. For parameter estimation, researchers need the sample size that give the required width of a confidence interval. In hypothesis testing, researchers require the sample size that gives a high probability of precisely identifying a hypothesis as true or false.

It was suggested by Clarke and Yuan (2006) that “sensitivity analyses should be used to ensure the sample sizes obtained from any one method are robust against deviations of the prior, likelihood and loss function (if one exists).” Sensitivity analysis is a method used to assess the impact that different values of an independent variable will have on a dependent variable or the impact that the change in some assumptions will have on the conclusions (Thabane *et al.* (2013); Schneeweiss (2006); Viel *et al.* (2007)).

In our research, we aim to use the Bayesian method to examine the impact of a different value of a parameter of the prior distribution on the posterior distribution. Thus some ideas in Clarke and Yuan (2006) are relevant to our research. On the other hand, the results in the paper are not relevant, because the sample size obtained in the paper is fixed. We plan to use interim analyses so we can halt the trial when we have enough evidence that one arm is better. In our research, the sample size is not fixed.

Brutti *et al.* (2008) applied a robust Bayesian approach to sample size determination. According to Brutti *et al.* (2008) and Greenhouse and Wassermann (1995), the robust Bayesian approach is a method that considers a class of prior distributions instead of one prior distribution with specific values of the parameters. This approach studies the change in the posterior distribution when using different values of the parameters of the prior distribution. If the posterior distributions are similar to one another, it means that the design is robust. On the other hand, if the posterior distributions are substantially different, it means that the design is not robust. For hypothesis testing, the sample size determination (SSD) method depends on the test's power function, which is assessed under the alternative hypothesis (conditional power). Conditional power is based mainly on the design values. The design values are the values of the parameters that are set at the beginning of the trial; e.g., scenario 1 (Huang *et al.* (2009), $p_1 = 0.2, ...; \mu_1 = 4, ...$). Since the SSD method is based on the guessed values of the parameter, the resulting sample sizes are only locally optimal. Local optimality can be defined as a result that is optimal for specific values of the parameters so it is not globally optimal. In order to avoid local optimality, many authors have supported a Bayesian approach that models uncertainty on the design values.

Brutti *et al.* (2008) suggested that the sample size can be chosen by considering the predictive distribution of the posterior probability. The predictive distribution is the distribution of future data (prediction) conditional on the observed data.

The sample size obtained in Brutti *et al.* (2008) is fixed. Thus the results of the paper are not directly relevant to our research. However, the robust Bayesian approach is very relevant, because in the research we plan to use a class of prior distributions and then update it after the information from the trial becomes available to obtain the posterior distributions.

The Bayesian approach was also used by Sylvester (1988) who used a decision theory approach for the Phase II design by considering:

- *a priori* information

  This information will be obtained by considering the efficacy of a new treatment or the response rates of other new treatments from previous trials conducted for the same disease;

- the costs that are incurred when treating a patient with a new treatment;

- the benefits or losses arising from the decisions made at the end of this phase.

This decisional approach gives a formal determination for the sample sizes used frequently in the Phase II study. I wrote an R program and verified the results of Sylvester (1988).

Sylvester (1988) is an interesting paper because it emphasises that the aim of clinical trials is to obtain correct conclusions and also to reduce the cost of the trial. Hence, decision theory should be used when conducting trials. In addition, the cost of the trial should be considered: not only the cost of the Phase II trial alone, but also the costs after a new treatment is given to a patient. As the optimal sample size calculated in the paper is used in a fixed-sample size clinical trial, the paper is not completely relevant to our research. However, the three dot points above are of importance in this research.

In our research, we focus not only on the Bayesian Method but also on survival analysis.

According to Kleinbaum and Klein (2005), the aim of survival analysis is to study the length of time after receiving a treatment until death. In many investigations, the death of all patients has not occurred by the end of the trial. At the end, we may therefore keep the information that the remaining subjects were still alive. However we do not know when they will die. This case is called censored data. In addition, if patients are lost to follow up during the course of study, the situation will be also considered as censored data. Brooks (1982) examined the information loss when data on lifetimes were censored, and considered this data in both reliability and survival studies.

Although in this research we focus on survival analysis, Brooks (1982) is not directly relevant. This is because we only use the censored survival times to update the posterior distribution. We are not concerned about the loss of information, as we have no way to avoid the data being censored.

In a real clinical trial, we may encounter complex situations. Lakatos (1988) developed a method of estimating sample sizes for the Log-Rank Statistic when the risk of the event of interest varies. He considered a situation that allows any pattern of survival including noncompliance, drop-in and loss to follow-up. In this situation, although the effect of the treatment was constant throughout the time, the hazard rate did not stay constant. The hazard function is the probability that an event of interest (e.g. death) occurs in the next instant, given survival to time t (Kleinbaum and Klein, 2005). For example, according to Gail (1985), the hazard rates of “drop-in” patients can be supposed to be 5 percent per year.

Lakatos (1988) used a Markov model to calculate the sample size. In this model, patients were assumed to be in one of four states, namely loss to follow-up, the event occurs, ‘Active complier’ and ‘Active noncomplier’. The Log Rank

Statistic was used in this paper because he estimated the hazard rates of non-compliance, loss to follow-up and drop-ins in both the treatment and control groups.

Lakatos (1988) said that "administratively censored" observations mean that an event of interest does not happen in the length of the trial; the trial for that patient is stopped for administrative reasons.

In many trials, participants are recruited to a clinical trial for the required duration, and thus the length of follow-up time differs from one individual to another. This is called staggered entry, or extended accrual (Lakatos, 1988, Shih, 1995).

We wrote an R program in an attempt to follow Lakatos's procedure. However, some results that it gave were different from those in Lakatos (1988). My supervisor sent an email of enquiry to the author. Lakatos sent us a MATLAB program that does the calculations. We ran the MATLAB program and obtained the same results as our R program gave. We therefore conclude that there were errors in the numerical results in the Lakatos (1988) paper.

The situation mentioned in this paper is different from the situation in our research. Consequently, in our research, Lakatos (1988)'s situation will not be directly used. This is because we will carry out the simulation in accord with Huang et al (2009)'s situation. However, some content of Lakatos (1988) in the paper is relevant to our research because our patients have staggered entry, i.e. they do not arrive simultaneously. In addition, censored survival times are used to update the posterior distribution. We can use the definition of staggered entry and administratively censored observations in our research.

Lakatos (2002) considered the design of group sequential trials. According to Pocock (1977), a group sequential design is an adaptive design that allows for interim analyses. In this design, a group of patients are enrolled sequentially.

Then an evaluation is performed at periodic intervals. Group sequential trials are widely used in Phase II clinical trials. Lakatos introduced a Markov model in Lakatos (1988) for the group sequential design. In addition, this model was used to calculate the sample size for each interim analysis.

This paper is important because this design can be used in Phase II clinical trials. In addition our patients are recruited sequentially, so the design used is a sequential design. We are updating our information before the trial ends, hence we would also be conducting interim analyses. However, we are not looking at groups of patients, so that a group sequential design is not used in this research. We absolutely do not consider groups of patients.

As stated above, one method that will be used in our research is interim analyses. By conducting interim analyses, early stopping for efficacy will be relevant. We determine every week whether the trial should be stopped early. Therefore a hypothesis test will be performed every week. The trial will be stopped early if arm A (or arm B) is selected as a superior treatment. That is, the posterior probability of assigning the next patient who is enrolling to treatment A, p_A, has fallen in the critical region. However, in each interim analysis we will carry out the global hypothesis test. Chang (2008, pp. 54 - 55) considered an interim analysis of K stages. In each stage, he considered a local hypothesis test. Our trial is not divided into K stages. We therefore cannot use the method of Chang.

Another example of adaptive design is the response-adaptive randomisation. Bather (1981) considered the multi-armed bandit problem, which is a statistical decision model. Its purpose is to determine a rule for assigning a treatment to a patient by depending on former outcomes which are successes or failures. An effective rule should identify the best treatment and treat each patient as effectively as possible. One method with this strategy is ‘play-the-winner’, which is referred to as a response-adaptive randomisation. This method can be used to compare two treatments with dichotomous outcomes. The basic idea can be described as follows: the first subject is assigned to a treatment by using equal randomisation. Then the next patient is allocated to the treatment with the higher probability of success (Rosenberger and Lachin, 1993). Therefore the probability of assigning a patient depends upon the probability of success for each treatment.

In our research, we will focus on comparing two treatments and use the response-adaptive randomisation. However, our strategy is not the play-the-winner, because our research will be carried out to compare two treatments for a continuous response (the progression-free survival time).

Rosenberger and Hu (2004) suggested that currently the main concerns when comparing clinical trials are the power of the test, sample size, rate of treatment failures, etc. Bandyopadhyay and Bhattacharya (2006) offered a response adaptive design for comparing two treatments with a continuous response. In this paper, the response variables were assumed to be normally distributed. The treatment allocation rule was considered from two aspects:

- minimizing the rate of treatment failures;

- the power of the test.

In our research we will use the response-adaptive randomisation. We also aim to compare two treatments with a continuous response. However, Bandyopadhyay and Bhattacharya (2006) is not directly relevant to our research since it did not consider the prior distributions for various parameters as in Huang *et al.* (2009).

1.6 Huang *et al.* (2009)

This section provides more detail of Huang *et al.* (2009)(hereafter referred to as 'HNL'). As described in the previous section, the designs of HNL are adaptive designs which combine response-adaptive randomisation and interim analyses.

In these designs, both survival and short-term patient response are used as primary endpoints. Often, survival is used as the primary endpoint because it is the final aim of the medical treatment. However, the disadvantages of a survival endpoint are that it requires a lot of time to elapse and it causes some difficulty when using response-adaptive randomisation.

In response-adaptive randomization, the probability of assigning a treatment to the next patient is based on the response from the previous patients. That is, more patients will be assigned to the demonstrably better arm. To assign a treatment to the subsequent patient, the results so far are required immediately. This indicates why the survival endpoint causes some difficulty when using response-adaptive randomisation.

Hence, both survival and short-term patient response are used as the primary endpoints.

According to Thall and Wathen (2005), in a cancer trial, patient response is commonly classified into four categories depending upon the outcome of the patient after completing treatment:

- progressive disease, or death,
- stable disease,
- partial remission (PR),
- complete remission (CR).

In HNL, short-term response was also considered in the four categories described above. Then a mean progression-free survival time was assigned to each category.

According to Green *et al.* (2008) and HNL, the progression-free survival time can be defined as the length of time from receiving a treatment until the first event occurs. This event could be resistance, progress of disease, deterioration, or death.

1.6.1 Theory

In this section, theory used in HNL will be explained. Then we will show how to obtain the posterior distributions of $\mu_{x,k}$ and $p_{x,k}$ where $\mu_{x,k}$ is the mean progression-free survival time of the kth category in arm x and $p_{x,k}$ is the probability of a patient in arm x occupying the kth category of a short-term response, where $p_{x,k} > 0$ for each k, and $p_{x,1} + p_{x,2} + p_{x,3} + p_{x,4} = 1$.

Note that we adopt the notation of HNL in many cases.

In an HNL design, the Bayesian model is used to connect the short-term response with survival response. This is because the model is established by using the information from the short-term response to update the prior distribution of the probabilities of being in the four categories and also to update the prior distribution of the mean progression-free survival time of each category. It can be also used to predict the long-term survival of patients.

Here we will illustrate how to obtain the posterior distributions on $\mu_{x,k}$.

Let x represent the treatment arm ($x = a$ for treatment A, $x = b$ for treatment B).

Let $T^k_{x,i}$ be the progression-free survival time of participant i in arm x if this patient occupies the kth category. $T^k_{x,i}$ is assumed to have an exponential distribution with rate $\lambda_{x,k}$.

Let $t^{(k)}_{x,i}$ be the observed or censored survival time of patient i in arm x in the kth category. The likelihood of the $t^{(k)}_{x,i}$ will be provided below.

In Bayesian probability theory, if a chosen prior distribution has a suitable form, then the posterior distribution will be of the same family as the prior distribution. The choice of the family is based on the likelihood. Prior and posterior distributions chosen to achieve this are said to be conjugate, and the prior distribution is called the conjugate prior distribution of the likelihood. The conjugate prior for the parameter $\lambda_{x,k}$ of the exponential distribution is the Gamma distribution.

This research focuses on the mean time between events, $\mu_{x,k} \equiv 1/\lambda_{x,k}$, so that the conjugate prior for the parameter $(1/\lambda_{x,k})$ of the exponential distribution, that is the Inverse Gamma distribution, is used.

Let us consider the prior distribution on $\mu_{x,k}$.

Initially, $\mu_{x,k} \equiv 1/\lambda_{x,k}$ is assumed to have an Inverse Gamma $(\alpha_{x,k}, \beta_{x,k})$ distribution, with shape parameter $\alpha_{x,k}$ and scale parameter $\beta_{x,k}$. Thus, the probability density function of the Inverse Gamma distribution is

$$f(\mu_{x,k}) = f(1/\lambda_{x,k}) = \frac{\beta^{\alpha_{x,k}}}{\Gamma(\alpha_{x,k})}(\frac{1}{\lambda_{x,k}})^{-\alpha_{x,k}-1}e^{-\beta_{x,k}\lambda_{x,k}} \text{ for } \lambda_{x,k} > 0,$$

where $\alpha_{x,k}$ and $\beta_{x,k}$ are initial simulation parameters that will be updated as the clinical trial progresses. Now we are going to derive the likelihood of a single observation $t_{x,i}^{(k)}$.

Let $\delta_{x,i}^{(k)}$ denote a dummy variable associated with $t_{x,i}^{(k)}$ where $\delta_{x,i}^{(k)} = 0$ for censored time and $\delta_{x,i}^{(k)} = 1$ for observed time.

For observed times,

$$f(t_{x,i}^{(k)}) = \lambda_{x,k}e^{-\lambda_{x,k}t_{x,i}^{(k)}}$$

for $t_{x,i}^{(k)} \geq 0$. Since the censored survival times are longer than observed times, the probability of the censored times is

$$P(T_{x,i}^{(k)} > t_{x,i}^{(k)}) = 1 - P(T_{x,i}^{(k)} \leq t_{x,i}^{(k)}) = e^{-\lambda t_{x,i}^{(k)}}.$$

The likelihood of a single observation $t_{x,i}^{(k)}$ can be expressed as

$$(\lambda_{x,k} \exp{(-\lambda_{x,k} t_{x,i}^{(k)})})^{\delta_{x,i}^{(k)}} (\exp{(-\lambda_{x,k} t_{x,i}^{(k)})})^{1-\delta_{x,i}^{(k)}} = \lambda_{x,k}^{\delta_{x,i}^{(k)}} \exp{(-\lambda_{x,k} t_{x,i}^{(k)})}$$

for $t_{x,i}^{(k)} \geq 0$ and

$$\delta_{x,i}^{(k)} = \begin{cases} 0 & \text{for censored time,} \\ 1 & \text{for observed time.} \end{cases}$$

For a vector of independent and identically distributed (i.i.d.) observations $(t_{x,1}^{(k)}, ..., t_{x,n}^{(k)})$, the likelihood function is given by

$$\begin{aligned} & \prod_{i=1}^{n_{x,k}} \lambda_{x,k}^{\delta_{x,i}^{(k)}} \exp{(-\lambda_{x,k} t_{x,i}^{(k)})} \\ = & \lambda_{x,k}^{\sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}} \exp{(-\lambda_{x,k} \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})} \end{aligned}$$

By Bayes' theorem, the posterior distribution satisfies

$$\begin{aligned} f(\mu_{x,k} \mid t_{x,i}^{(k)}) & \propto \frac{\beta^{\alpha_{x,k}}}{\Gamma(\alpha_{x,k})} (\frac{1}{\lambda_{x,k}})^{-\alpha_{x,k}-1} \exp{(-\beta_{x,k}\lambda_{x,k})} \lambda_{x,k}^{\sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}} \exp{(-\lambda_{x,k} \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})} \\ & \propto \frac{\beta^{\alpha_{x,k}}}{\Gamma(\alpha_{x,k})} (\frac{1}{\lambda_{x,k}})^{-(\alpha_{x,k}+\sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)})-1} \exp{[-(\beta_{x,k} + \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})\lambda_{x,k}]}. \end{aligned}$$

Therefore, the posterior distribution of $\mu_{x,k}$ is IG$(\alpha'_{x,k} = \alpha_{x,k} + \sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}, \beta'_{x,k} = \beta_{x,k} + \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})$.

Now we give derivations of the posterior distributions of $p_{x,k}$.

Let $S_{x,k,i}$ denote the short-term response in the kth category for patient i in treatment x. If this patient occupies in the kth category, $S_{x,k,i} = 1$ and $S_{x,j,i} = 0$ for $1 \leq j \leq 4, j \neq k$. Suppose that the vectors $(S_{x,1,i}, ..., S_{x,4,i})$ are i.i.d. across

$i = 1, ..., n_x$ and have a multinomial $(1, p_{x,1}, ..., p_{x,4})$ distribution. Thus,

$$f(S_{x,1,i}, S_{x,2,i}, S_{x,3,i}, S_{x,4,i}) = (p_{x,1})^{S_{x,1,i}} \times (p_{x,2})^{S_{x,2,i}} \times (p_{x,3})^{S_{x,3,i}} \times (p_{x,4})^{S_{x,4,i}}$$

where $S_{x,1,i} + S_{x,2,i} + S_{x,3,i} + S_{x,4,i} = 1$.

Suppose we observe the vectors $(S_{x,1,1}, ..., S_{x,4,1}), ..., (S_{x,1,n_x}, ..., S_{x,4,n_x})$. Then the likelihood function is given by

$$f(n_{x,1}, ..., n_{x,4}; p_{x,1}, ..., p_{x,4}) = \frac{n_x!}{n_{x,1}!n_{x,2}!n_{x,3}!n_{x,4}!}(p_{x,1})^{n_{x,1}}(p_{x,2})^{n_{x,2}}(p_{x,3})^{n_{x,3}}(p_{x,4})^{n_{x,4}}$$

where n_x is the total number of the patients allocated to arm x, $n_{x,k}$ is the number of patients in arm x falling in the kth category of a short-term response, and $n_{x,1} + n_{x,2} + n_{x,3} + n_{x,4} = n_x$.

The conjugate prior for the parameters $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ of the multinomial distribution is the Dirichlet distribution.
Suppose that $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ has a Dirichlet $(\gamma_{x,1}, \gamma_{x,2}, \gamma_{x,3}, \gamma_{x,4})$ distribution. Therefore, the probability density function of $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ is

$$f(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4}) = \frac{1}{\mathrm{B}(\gamma_{x,1}, \gamma_{x,2}, \gamma_{x,3}, \gamma_{x,4})}(p_{x,1})^{\gamma_{x,1}-1}(p_{x,2})^{\gamma_{x,2}-1}(p_{x,3})^{\gamma_{x,3}-1}(p_{x,4})^{\gamma_{x,4}-1}$$

when $\gamma_{x,1}, \gamma_{x,2}, \gamma_{x,3}, \gamma_{x,4} > 0$. By Bayes' theorem, the posterior distribution of $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ satisfies

$$\begin{aligned} f(p_{x,1}, ..., p_{x,4} \mid n_{x,1}, ..., n_{x,4}) \propto & \frac{n_x!}{n_{x,1}!n_{x,2}!n_{x,3}!n_{x,4}!\mathrm{B}(\gamma_{x,1}, \gamma_{x,2}, \gamma_{x,3}, \gamma_{x,4})} \\ & \times (p_{x,1})^{(\gamma_{x,1}+n_{x,1})-1}(p_{x,2})^{(\gamma_{x,2}+n_{x,2})-1}(p_{x,3})^{(\gamma_{x,3}+n_{x,3})-1}(p_{x,4})^{(\gamma_{x,4}+n_{x,4})-1} \end{aligned}$$

Hence, the posterior distribution of $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ is Dir $(\gamma_{x,1} + n_{x,1}, \gamma_{x,2} + n_{x,2}, \gamma_{x,3} + n_{x,3}, \gamma_{x,4} + n_{x,4})$.

From the derivations above, we see that the statistical theory underlying the proposed clinical trials is mathematically complicated. Consequently most evaluations of the proposed trials are done by simulation, which is more feasible. The detail of the simulation will be described in the following chapter.

After the information from a patient becomes available, Bayes' theorem is used to update the prior distributions of $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ and $\mu_{x,k}$ to evaluate the posterior probability $p_A = Pr(\mu_a > \mu_b \mid data)$ where μ_x is the mean progression-free survival time for arm x, and is defined as

$$\mu_x = \sum_{k=1}^{4} p_{x,k}\mu_{x,k}$$

for $x = a, b$.

This posterior probability, p_A, is applied to a response-adaptive randomisation procedure and provides a criterion for an interim analysis.

In HNL, three outcomes were considered. The authors decided that treatment A (or B) would be chosen as the better treatment and the trial would be terminated if $p > p_U$ (or $p < p_L$) for some constants p_L and p_U. Additionally, the trial would not reach a conclusion if the maximum number of patients had been recruited and, if, at the end of the study, neither treatment A nor treatment B had been chosen as superior.

Finally HNL compared the results of adaptive and common designs. A 'common design' is a design that uses response-adaptive randomisation in a manner similar to the adaptive design. However, a 'common design' does not use the information from short-term patient response. This is different from the adaptive design. The 'common design' was introduced by HNL because they aimed to show that, by using short-term response, the adaptive design gave higher power than the 'common design.' They found that the adaptive design requires a smaller number of patients, and it allows more patients to be assigned to the better treatment than the 'common design.'

1.7 Overview of this thesis

The aim of this thesis is to generalise extensively the adaptive designs of HNL. Hence, we will develop this design through two aspects, (1) the recruitment regime and (2) the randomisation procedure, by considering the response of the previous patients and some prognostic factors which may be important when considering a cancer trial. Additionally, we will consider whether we can use better procedures that incorporate prognostic factors. We will also investigate important criteria for evaluating and comparing the HNL designs with competing designs. Then we will show an example of using these criteria for assessing and comparing the designs.

Chapter 2 will investigate the adaptive method of HNL. Our main concern is to consider a different recruitment regime for this method. In this enrolment regime, the accrual rate will be changed from exactly one patient per week to an average of one per week. In reality, an accrual schedule of exactly one patient per week rarely occurs. An investigation into whether this more realistic scenario affects the results obtained by HNL from simulation will be carried out.

When designing a clinical trial, researchers need to evaluate and compare a proposed design with other designs in order to ensure that the proposed design is effective. This leads to Chapter 3. In this chapter, the principal criteria for evaluating and comparing designs will be addressed and employed. In particular, we will focus on several criteria: the Operating Characteristic Curve, and the design characteristics. Then we will show how to apply these criteria to evaluate and compare designs.

Chapter 4 will extend the HNL design to a design that is applicable to a more realistic situation, as the HNL design does not have an appropriate randomisation procedure.

In the HNL design, the response adaptive randomization is used. Hence, the assignment of a treatment to a new patient is based only upon the response of the previous patients. However, it does not consider the possibility that important prognostic factors might influence the effect of the treatments. In order to fill this gap in the HNL design, a subsequent patient will be allocated to a treatment by considering not only the response of the previous patients but also the prognostic factors. Then the extension of the HNL design will be evaluated by employing the criteria in Chapter 4. The simulation results obtained from four designs will be provided and compared.

Lastly, Chapter 5 will contain the conclusions from the research described in the preceding chapters and a brief discussion of future research.

Chapter 2

Investigation of the Huang *et al.* (2009) method

The aim of this chapter is to investigate the adaptive method of HNL. One step will be to consider a different enrolment regime for this method. In this regime, we will change the arrival rate from exactly one patient per week to an average of one per week. We will examine whether this more realistic scenario affects the results obtained from simulation.

We begin by investigating the HNL program. The next step is to compare the results obtained using the HNL program with their published results. Then an investigation of the adaptive method of HNL is performed. Finally, we will investigate the difference between the results obtained from an average of one arrival per week and the results obtained from exactly one arrival per week.

2.1 Huang *et al.* (2009) program

We examined the HNL program and tried to understand the underlying algorithm. In doing this, some discrepancies and errors were found. The program from which the HNL results were obtained is available as Ning (2009). The program was downloaded from the MD Anderson Cancer Center website ($https://biostatistics.mdanderson.org/$ $SoftwareDownload/SingleSoftware.aspx?Software\_Id = 82$). We then checked

Ning (2009) by comparing the results obtained using this program with published results in HNL.

2.1.1 Aspects of Ning (2009)

In HNL, a trial was conducted to compare two treatments (A and B) for a continuous response (the progression-free survival time). Recall that, in HNL, the vectors $(S_{x,1,i}, ..., S_{x,4,i})$ were assumed to have a multinomial $(1, p_{x,1}, ..., p_{x,4})$ distribution where $S_{x,k,i}$ is the short-term response in the kth category for patient i in arm x. If this patient belongs in the kth category, $S_{x,k,i} = 1$ and $S_{x,j,i} = 0$ for $1 \leq j \leq 4, j \neq k$. In HNL, $T_{x,i}^{(k)}$ was defined as the progression-free survival time of patient i if she/he is in category k and arm x. $T_{x,i}^{(k)}$ was assumed to have an exponential distribution with rate $\lambda_{x,k}$ where $\lambda_{x,k} = 1/\mu_{x,k}$.

Ning (2009) began by generating treatments for the n_0 initial patients before adaptive randomization commenced. These treatments were drawn from a Bernoulli (1, 0.5) distribution. After these patients were allocated to treatments A or B, for each patient, the category variable was simulated. Then $T_{x,i}^{(k)}$ was simulated.

In Ning (2009), there were two sets of parameters $\{p_{x,k}$ and $\mu_{x,k}\}$ with the same labels, one set for data generation and another for estimation procedures.

Data generation procedure

In the HNL data generation procedure, $p_{x,k}$ and $\mu_{x,k}$ were used to generate the category variable and the $T_{x,i}^{(k)}$.

It should be noted that, in this thesis, we will introduced the symbols $\pi_{x,k}$ and $\nu_{x,k}$ to distinguish between the values of parameters $p_{x,k}$ and $\mu_{x,k}$ for data generation and the values for probability estimation.

To avoid confusion of the $p_{x,k}$ for the estimation procedure with the $p_{x,k}$ used in the simulations, denote by $\pi_{x,k}$ the value that was given to this parameter for each scenario for the simulations (e.g. under Scenario 1, $\pi_{x,1} = 0.2, \pi_{x,2} = 0.4, \pi_{x,3} = 0.1, \pi_{x,4} = 0.3$). Thus $(S_{x,1,i}, S_{x,2,i}, S_{x,3,i}, S_{x,4,i})$ for all patients was drawn from a multinomial $(1, \pi_{x,1}, \pi_{x,2}, \pi_{x,3}, \pi_{x,4})$ distribution.

Additionally, denote the mean progression-free survival time for the simulations by $\nu_{x,k}$ (e.g. under Scenario 1, $\nu_{x,1} = 4, ..., \nu_{x,4} = 110$). Therefore, the $T_{x,i}^{(k)}$ for all patients were drawn from an exponential distribution with $\lambda_{x,k} = 1/\nu_{x,k}$.

In Scenario 1, in order to evaluate this design by looking at the Type I error rate, the values of $\pi_{x,k}$ and $\nu_{x,k}$ in the two arms are identical. In this scenario, the hypotheses are

$H_0 : \mu_a - \mu_b = 0$

$H_1 : \mu_a - \mu_b \neq 0.$

In contrast, in Scenarios 2 and 3, arm B is assumed to be a superior treatment. The hypotheses will be

$\mathrm{H}_0 : \mu_a \geq \mu_b;$

$\mathrm{H}_1 : \mu_a < \mu_b.$

Due to the assumption in Scenarios 2 and 3, HNL can evaluate the effectiveness of this design by looking at power. An effective design should have high power. That is, it should identify the correct superior treatment frequently.

It should be noted that, in a given set of simulations, $\pi_{x,k}$ and $\nu_{x,k}$ were kept constant. This is because in this procedure, $p_{x,k} = \pi_{x,k}$ and $\mu_{x,k} = \nu_{x,k}$. Due to this, $(S_{x,1,i}, S_{x,2,i}, S_{x,3,i}, S_{x,4,i})$ and $T_{x,i}$ are i.i.d. across $i = 1, ..., n_x$.

Estimation procedure

In this procedure, Bayes' theorem was used to update the prior distributions of $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ and $(\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4})$ after the n_0 initial patients had entered the trial (i.e. before adaptive randomization). Then the program computed the mean progression-free survival time $\mu_x = \sum_{k=1}^{4} p_{x,k}\mu_{x,k}$ and evaluated the posterior probability of assigning the next patient who was enrolling to treatment A, $p_A = Pr(\mu_a > \mu_b \mid data)$. Ning (2009) generated the values of $\mu_{x,k}$ from an IG($\alpha'_{x,k} = \alpha_{x,k} + \sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}, \beta'_{x,k} = \beta_{x,k} + \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)}$) distribution. In addition, the values of $p_{x,k}$ were generated using $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4}) \sim$ Dir$(\gamma_{x,1} + n_{x,1}, \gamma_{x,2} + n_{x,2}, \gamma_{x,3} + n_{x,3}, \gamma_{x,4} + n_{x,4})$.

For those patients who entered the trial after adaptive randomization commenced, their treatments were drawn from a Bernoulli (1, 1 - p_A) distribution where $x = 0$ for arm A and $x = 1$ for arm B.

The posterior probability p_A is also used to determine which treatment is selected as the superior treatment. If the design is effective, it can identify the superior treatment correctly. In all scenarios, the initial values of $\gamma_{x,k}$, $\alpha_{x,k}$ and $\beta_{x,k}$ were identical for arms A and B. It was assumed by HNL that $\gamma_{x,k} = 0.5$ and $\alpha_{x,k} = 11$ for $k = 1, 2, 3, 4$ and $x = a, b$. It was assumed also that $\beta_{x,1} = 40, \beta_{x,2} = 300, \beta_{x,3} = 750$ and $\beta_{x,4} = 1100$ for $x = a, b$. This values were chosen by HNL "The amount of information in these prior distributions is approximately equal to that from 11 patients".

By assuming the parameter values in the previous paragraph, at the beginning of the trial the prior distributions of $(p_{x,1}, p_{x,2}, p_{x,3}, p_{x,4})$ and $(\mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \mu_{x,4})$ were the same for arms A and B. However, as the trial progressed, the prior distributions were updated. The accumulated information is expected to help the clinical trial determine whether there is a difference between the two treatments.

In the situation described above, Ning (2009) made a decision as to whether the trial should be terminated early. If n_0 equalled 1, in the second week the prior distributions of $(p_{x,1}, ..., p_{x,4})$ and $(\mu_{x,1}, ..., \mu_{x,4})$ were updated. The posterior probability of assigning patients to arm A was also updated and used to determine whether to stop the trial early. This procedure was done each week until 120 patients had been admitted or the trial was terminated.

In Ning (2009), *after all patients had entered the trial*, the authors did not update the prior distributions of $(p_{x,1}, ..., p_{x,4})$ and $(\mu_{x,1}, ..., \mu_{x,4})$, nor did they evaluate the posterior probability p_A separately during the follow-up period. Therefore, the trial could not be stopped early once the enrolment period had ended. However, if we were to modify the original program, we could update the prior distributions of $(p_{x,1}, ..., p_{x,4})$ and $(\mu_{x,1}, ..., \mu_{x,4})$ each week not only during the recruitment period but also during the follow-up period. Therefore, the evaluation of the posterior probability would be done every week.

Normally, either arm could be superior. However, for simplicity, arm B was assumed to be the superior treatment if arms A and B were not identical.

We note that, unlike $\pi_{x,k}$ and $\nu_{x,k}$ in the data generation procedure, $p_{x,k}$ and $\mu_{x,k}$ in the probability estimation procedure were drawn from the Dir $(\gamma_{x,k} + n_{x,k})$ and IG $(\alpha_{x,k} + \sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}, \beta_{x,k} + \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})$ distributions respectively.

2.1.2 Errors in Ning (2009)

Ning (2009) requested input in months. However in HNL, they looked at patients per week. In the description of the parameters, ‘atime’ should be ‘addtime’ and the description of ‘alpha’ was incomplete. In this program, ‘alpha’ represents the value of $\gamma_{x,k}$ which is a parameter for the Dirichlet distribution. In requesting input to the program, the scale parameter for the Inverse-gamma distribution was described as the mean survival time multiplied by the number of patients. However it was actually the mean survival time multiplied by (the number of patients minus 1). The variable ‘fstop’ was initialised, and then updated each time an arm was selected during the n_0 weeks, but there was no further use of this variable. It would seem that it could be completely removed from the program without affecting it.

2.1.3 The results from Ning (2009)

In this section, the results obtained using Ning (2009) will be compared with the published results of HNL.

In HNL, for a given set of design parameters, a total of 5,000 simulations was carried out to evaluate the performance of the HNL and common designs. The performances of these designs were compared by using four design characteristics: the probability of Type I error, the power of the test, the average number of patients allocated to each arm, and the average number of patients in the trial. Details of these design characteristics will be provided in the next chapter.

Table 2.1 and Table 2.2 show the results using Ning (2009) for the HNL design, compared with the published results of HNL when using $p_U = 0.975$, the maximum sample size (N_{Max}) $= 120$ and $n_0 = 1$ and 30 respectively. Likewise, Table 2.3 and Table 2.4 show the results obtained using Ning (2009) for the common design, compared with the published results of HNL when using $p_U = 0.993$, $N_{Max} = 120$ and $n_0 = 1$ and 30 respectively.

The results in Table 2.1 illustrate that we matched the published results of HNL in Scenarios 1 and 2. However, we did not match their results in Scenario 3. This raised a question mark over their results in Scenario 3.

The results displayed in Table 2.2 show that the results obtained using Ning (2009) were different from the published results of HNL. It can be seen that the probability of type I error obtained using Ning (2009) is nearly half the value of the published results. On the other hand, the results shown in Tables 2.3 and 2.4 illustrate that, in Scenario 1, the results obtained using Ning (2009) shown in these tables were similar to the published results of HNL. We did not match their results for the other Scenarios.

We wondered why we could not match some published results even though we used their program and their parameters. My supervisor sent an email of enquiry

Table 2.1: Comparison of the results for the HNL design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.975$, $N_{Max} = 120$ and $n_0 = 1$

Scenario	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	Source	Average	Probability of selection as the superior arm
1	A	0.2	0.4	0.1	0.3	4	30	75	110	A	58.0964	0.0430
										B	58	0.046
										C	58.3226	0.0408
	B	0.2	0.4	0.1	0.3	4	30	75	110	A	57.7900	0.0398
										B	58	0.048
										C	57.6034	0.0424
	total									A	115.8864	
										B	116	
										C	115.9260	
2	A	0.2	0.4	0.1	0.3	4	30	75	110	A	16.0056	0.0002
										B	16	0.0002
										C	15.7762	0.0008
	B	0.1	0.1	0.2	0.6	4	30	75	110	A	69.5080	0.5904
										B	71	0.59
										C	69.4464	0.6002
	total									A	85.5136	
										B	87	
										C	85.2226	
3	A	0.2	0.4	0.1	0.3	4	30	75	110	A	11.6862	0.0002
										B	11	0.0002
										C	11.9036	0
	B	0.1	0.1	0.2	0.6	4	30	75	110	A	56.6884	0.9028
										B	51	0.976
										C	57.7644	0.9020
	total									A	68.3746	
										B	62	
										C	69.6680	

Let source A represent the results using Ning (2009) with seed 1234; source B represent the published results of HNL; and source C represent the results using Ning (2009) with seed 5678.

Table 2.2: Comparison of the results for the HNL design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.975$, $N_{Max} = 120$ and $n_0 = 30$

Scenario	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	Source	Average	Probability of selection as the superior arm
1	A	0.2	0.4	0.1	0.3	4	30	75	110	A	59.2334	0.0278
										B	58	0.048
										C	58.6892	0.0288
	B	0.1	0.1	0.2	0.6	4	30	75	110	A	58.4800	0.0228
										B	56	0.054
										C	58.9704	0.0250
	total									A	117.7134	
										B	114	
										C	115.7806	
2	A	0.2	0.4	0.1	0.3	4	30	75	110	A	24.3782	0
										B	23	0.0002
										C	24.0526	0
	B	0.1	0.1	0.2	0.6	4	30	75	110	A	66.5040	0.5758
										B	52	0.704
										C	66.5040	0.5758
	total									A	90.8822	
										B	75	
										C	90.5566	
3	A	0.2	0.4	0.1	0.3	4	30	75	110	A	20.5516	0
										B	20	0.0002
										C	20.4110	0
	B	0.1	0.1	0.2	0.6	4	30	75	110	A	54.8228	0.8930
										B	40	0.918
										C	54.8228	0.8906
	total									A	75.3744	
										B	60	
										C	82.6024	

Let source A represent the results using Ning (2009) with seed 1234; source B represent the published results of HNL; and source C represent the results using Ning (2009) with seed 5678.

Table 2.3: Comparison of the results for the common design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.993$, $N_{Max} = 120$ and $n_0 = 1$

Scenario	Arm	μ_x	Source	Average	Probability of selection as the superior arm
1	A	60	A	57.0356	0.0490
			B	59	0.046
			C	57.8096	0.0414
	B	60	A	57.8430	0.0458
			B	59	0.045
			C	57.2428	0.0516
	total		A	114.8786	
			B	118	
			C	115.0524	
2	A	55	A	32.2380	0.0056
			B	26	0.002
			C	32.7410	0.0044
	B	85	A	72.0948	0.3920
			B	77	0.429
			C	71.6232	0.4010
	total		A	104.3328	
			B	103	
			C	104.3642	
3	A	55	A	20.7580	0.0014
			B	21	0.001
			C	20.9422	0.0012
	B	127	A	63.2780	0.8322
			B	72	0.648
			C	62.5738	0.8456
	total		A	84.0360	
			B	93	
			C	83.5160	

Let source A represent the results using Ning (2009) with seed 1234; source B represent the published results of HNL; and source C represent the results using Ning (2009) with seed 5678.

Table 2.4: Comparison of the results for the common design obtained using Ning (2009) with the published results of HNL when using $p_U = 0.993$, $N_{Max} = 120$ and $n_0 = 30$

Scenario	Arm	μ_x	Source	Average	Probability of selection as the superior arm
1	A	60	A	57.5362	0.0428
			B	58	0.047
			C	58.0710	0.0428
	B	60	A	58.2444	0.0420
			B	58	0.047
			C	57.6376	0.0434
	total		A	117.6596	
			B	116	
			C	115.7086	
2	A	55	A	34.3592	0.0028
			B	29	0.004
			C	34.3592	0.0028
	B	85	A	71.0160	0.3966
			B	70	0.477
			C	70.2794	0.4078
	total		A	105.3752	
			B	99	
			C	104.6386	
3	A	55	A	23.2552	0.0006
			B	22	0.0004
			C	23.8484	0
	B	127	A	59.3472	0.8852
			B	58	0.832
			C	60.3108	0.8706
	total		A	82.6024	
			B	80	
			C	84.1592	

Let source A represent the results using Ning (2009) with seed 1234; source B represent the published results of HNL; and source C represent the results using Ning (2009) with seed 5678.

to the authors. Huang responded that

```
There may be a glitch in our simulations. For example,
the cut-off value we used might be 0.03, 0.035 etc, instead of 0.025.
We need to adjust this cut-off value so that the type I error
in scenario 1 is close to 10%, then use the same cut-off value
for all scenarios.
```

(received on 7/7/12)

The author said that their simulations may have a glitch. We therefore concluded that it was likely that some results of Ning (2009) might differ from their published results. However, one point where we disagreed with Ning (2009) was as follows:

Huang (personal communication sent on 7/7/12) said that, in their simulations, they may have used the value of the cut off that gives a Type I error rate of 10% in Scenario 1, instead of the value of the cut off mentioned in the paper. Furthermore, this cut-off value should be used for all scenarios. We do not know if they just used one cut-off value. However, if they did use this cut-off value in all scenarios, why in some scenarios were the results using Ning (2009) similar to their published results? For example, in Table 2.1, under scenario 1, it is clear that the probabilities of type I error obtained using Ning (2009) were similar to the published results. Additionally, under scenario 2, the power of the test obtained using Ning (2009) is in agreement with their results. Hence, we assumed that they used $p_U = 0.975$ and the same p_U as ours. Therefore, we disagreed with some of the results in their paper.

We felt that there were some discrepancies in some part of Table 2.2. This is because we obtained good agreement in scenario 1, for Table 2.1, Table 2.3 and Table 2.4.

For the common design, the reasons that in Scenarios 2 and 3 the results obtained using Ning (2009) were different from the published results of HNL are as follows:

Firstly, as mentioned above, there was a ‘glitch’ in the simulations in HNL.

In addition, the authors merely explained (HNL, p.1686) that, in Scenario 1, the survival times for subjects in the two treatments were assumed to follow exponential distributions with mean μ_A and μ_B respectively. They also assumed that the prior distributions of μ_A and μ_B had $\mathrm{IG}(\alpha, \beta)$ distributions with $\alpha = 2$ and $\beta = 60$. These parameter values provided ‘reasonably noninformative prior distributions’ with $\mu = \beta/(\alpha - 1) = 60$ which was approximately equal to the mean progression-free survival time under Scenario 1. So, for Scenario 1, we knew how they obtained the estimated values of μ_A and μ_B in the common design. They chose these values from the estimated values of μ_A and μ_B in the HNL design. For example, the values of μ_A and μ_B under Scenario 1 equal $0.2 \times 4 + 0.4 \times 30 + 0.1 \times 75 + 0.3 \times 110 = 53.3$. They then approximated the values of μ_A and μ_B by 60 (each).

For Scenarios 2 and 3, however, HNL did not specify the values of μ_A and μ_B that they used for the common design. Consequently, we needed to choose some values. We selected these values from the approximate values of the HNL parameters. For example, in Scenario 2, in the HNL design, μ_A and μ_B were 53.3 and 84.4 respectively. We approximated the values of the μ_A and μ_B for the common design in arm A by 55, and in arm B by 85. A similar approach was made in Scenario 3. Under Scenario 3, in the HNL design, μ_A and μ_B were 53.3 and 126.5. So the values of μ_A and μ_B for the common design were estimated as 55 and 127 respectively. This was our attempt to emulate the parameter used in HNL.

2.2 The recruitment regime

The principal objective of this chapter is to examine the recruitment regime of the HNL design. In HNL, patients came to a trial at an arrival rate of exactly one patient per week. This might be because HNL aimed to keep the process simple. However, we felt that this arrival rate was artificial. Consequently, we established a new enrolment regime by changing the accrual rate from *exactly* one patient per week to an *average* of one per week.

We used an exponential distribution with a mean of one patient per week to generate the waiting time before the next patient was recruited to the trial. This was done because the exponential distribution is a common distribution for arrival times. The exponential distribution also has the ‘memoryless’ property, that is $\mathrm{P}(T > s + t \mid T > t) = \mathrm{P}(T > s)$. See Asimow and Maxwell (2010, pp. 237) for proof of the ‘memoryless’ property. Regardless of the value of t, the time until the next arrival has the same distribution as the original exponential distribution for the interarrival time. If a week has elapsed since the previous patient, this does not make it more or less likely that another week will elapse before the next patient is enrolled. In addition, we chose the average of one per week since it was reasonable to compare with exactly one patient per week.

In this chapter, since patients came to a trial at an arrival rate of an average of one per week, the schedule for evaluating the p_A was different from that of HNL. In HNL, patients were enrolled in the trial at a rate of exactly one patient per week. After n_0 patients had entered the trial, the posterior probability p_A was evaluated every week. This was because the assignment of a new patient was based on the p_A. In contrast, we evaluated the p_A at the arrival time of a new patient instead of evaluating it every week.

Generally, in order to evaluate the p_A (see detail in the estimation procedure in Section 2.1.1), $\delta_{x,i}^{(k)}$ and $t_{x,i}^{(k)}$ were measured. Recall that $t_{x,i}^{(k)}$ is the observed or censored time of patient i if she/he occupies category k and arm x.

During the enrolment period, $t_{x,i}^{(k)}$ is given by

$$t_{x,i}^{(k)} = \min(\text{the current time} - i^{th}\text{arrival time}, T_{x,i}^{(k)}). \tag{2.1}$$

After all patients have entered the trial and no more calculations are done until the trial ends, the current time is replaced by the maximum duration of trial. The quantity $t_{x,i}^{(k)}$ is then given by

$$t_{x,i}^{(k)} = \min(\text{the maximum duration of trial} - i^{th}\text{arrival time}, T_{x,i}^{(k)}). \tag{2.2}$$

Recall that $\delta_{x,i}^{(k)}$ is a dummy variable associated with $t_{x,i}^{(k)}$ where $\delta_{x,i}^{(k)} = 0$ for censored time and $\delta_{x,i}^{(k)} = 1$ for observed time. During the enrolment period, if the finish time of patient i is less than or equal to the current time, then $\delta_{x,i}^{(k)} = 1$. Otherwise, $\delta_{x,i}^{(k)} = 0$. However, after all patients have entered the trial, the current time is replaced by the maximum duration of trial. Hence, if the finish time of patient i is less than or equal to the maximum duration of the trial, then $\delta_{x,i}^{(k)} = 1$. Otherwise, $\delta_{x,i}^{(k)} = 0$.

As discussed above, in the two recruitment regimes, the schedules for evaluating the p_A are different. Hence, in this chapter, the current time, the finish time of a patient i and the maximum duration of a trial used to calculate $\delta_{x,i}^{(k)}$ and $t_{x,i}^{(k)}$ are different from those in HNL as well.

In HNL, the current time is the arrival week of a new patient, which is the same as the order of all patients. For example, if patient i is enrolling in the trial, the current time will be the i^{th} week. Furthermore, the finish time of patient i can be given by $i + T_{x,i}^{(k)}$.

Let us consider $t_{x,i}^{(k)}$ and $\delta_{x,i}^{(k)}$ *when patient* $(i+1)$ *is enrolling.* In HNL, following (2.2), $t_{x,i}^{(k)} = \min((i+1) - i, T_{x,i}^{(k)})$. To calculate $\delta_{x,i}^{(k)}$, if $(i + T_{x,i}^{(k)})$ is less than or equal to $(i+1)$, $\delta_{x,i}^{(k)} = 1$. Otherwise, $\delta_{x,i}^{(k)} = 0$.

For example, let us consider $t_{x,2}^{(k)}$ and $\delta_{x,2}^{(k)}$. Suppose that patient 3 is arriving and $T_{x,2}^{(k)} = 5$. In HNL, $t_{x,2}^{(k)} = \min(3-2, 5) = 1$. Also, $\delta_{x,i}^{(k)} = 0$ because $2 + 5 = 7$ is greater than 3.

In contrast, in this chapter, the current time is the arrival time of a new patient. For instance, if patient i is recruited in the trial, the current time will be the arrival time of the i^{th} patient. Moreover, the finish time of patient i can be given by (the arrival time of the i^{th} patient + $T_{x,i}^{(k)}$).

Let us consider $t_{x,i}^{(k)}$ and $\delta_{x,i}^{(k)}$ when patient $(i+1)$ is arriving. In this chapter, following (2.2), $t_{x,i}^{(k)} = \min(((i+1)^{th}$ arrival time) - (i^{th} arrival time), $T_{x,i}^{(k)})$. To calculate $\delta_{x,i}^{(k)}$, if (the arrival time of i^{th} patient + $T_{x,i}^{(k)}$) is less than or equal to the arrival time of the $(i+1)^{th}$ patient, $\delta_{x,i}^{(k)} = 1$. Otherwise, $\delta_{x,i}^{(k)} = 0$.

For instance, let us consider $t_{x,2}^{(k)}$ and $\delta_{x,2}^{(k)}$. Suppose that patient 3 is enrolling and $T_{x,2}^{(k)} = 5$. Suppose also that the arrival times of the 2^{nd} and 3^{rd} patients are 2.5 and 4 respectively. Consequently, $t_{x,2}^{(k)} = \min(4 - 2.5, 5) = 1.5$. Also, $\delta_{x,i}^{(k)} = 0$ because $2.5 + 5 = 7.5$ is greater than 4.

In HNL, the maximum duration of the accrual period was 120 weeks because the maximum number of patients recruited was 120. However, in our research the maximum duration of the enrolment period was the arrival time of the 120^{th} patient. The maximum duration of the follow-up period in both HNL and our research was 40 weeks. Therefore, the maximum durations of trials in HNL and our research were 160 weeks and (the arrival time of the 120^{th} patient + 40 weeks) respectively.

Hence, in this chapter, our process is more complicated than that in HNL.

2.3 Comparison of the HNL designs under two arrival rates

Our main concern is to determine whether the results from the two recruitment regimes are different. However, if the results are different, we cannot conclude that the difference comes only from the two accrual patterns. Differences may have occurred from many sources of variation. Therefore, we require the criterion used to be a minimum important difference that could indicate that the two recruitment regimes lead to different results.

As mentioned in Section 2.1.3, in HNL four design characteristics were employed to compare the HNL and common designs. Similarly, in this section, these design characteristics are used to compare the results obtained from the two enrolment regimes.

There are two groups of design characteristics. The first group consistes of the probability of Type I error, and the power of the test. Hence, for this group, we tested whether the two arrival rates gave different probabilities (e.g. the probabilities of Type I error) or not. In contrast, the second group consisted of the average number of patients allocated to each arm and the average number of patients. In this group, they are both means, not probabilities. We therefore tested to determine whether the two arrival rates gave a different mean or not.

We decided that the difference between the recruitment regimes would be regarded as 'practically significant' (as distinct from 'statistically significant') if either

- the difference between the probabilities obtained from the two arrival rates is greater than 0.05, or
- the difference between the means under the two arrival rates is greater than 4.

These numbers were chosen because greater than 0.05 and 4 were thought to be enough to indicate that the differences were caused not only by the other sources of variation but also were caused by the accrual patterns.

In all cases we used 5,000 simulations.

2.3.1 Hypothesis Testing

We aim to test whether the difference between the probabilities obtained from the two accrual rates is greater than 0.05. Therefore, we propose to test the hypotheses

$\mathrm{H}_0 : |p_1 - p_2| \leq 0.05$

$\mathrm{H}_1 : |p_1 - p_2| > 0.05$

where p_1 represents the probability when the accrual rate of patients is exactly one patient per week and p_2 represents the probability when the accrual rate of patients is an average of one per week.

It can be seen that $\mathrm{H}_0 : |p_1 - p_2| \leq 0.05$ means $-0.05 \leq p_1 - p_2 \leq 0.05$. Thus we could test the hypothesis at a level of 5% by calculating a 95% confidence interval (CI) for $p_1 - p_2$. Only if the CI lies completely outside (-0.05, 0.05) do we reject H_0. Otherwise, we retain H_0.

The quantities p_1 and p_2 are independent because both are randomly generated from different trials. In addition $np_i \geq 5$ and $n(1 - p_i) = nq_i \geq 5$ for $i = 1, 2$. As a result, we use the normal distribution as an approximation of the binomial distribution.

A 95% CI estimate for the difference $p_1 - p_2$ is:

$$\left((\hat{p_1} - \hat{p_2}) - Z_{0.025}\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}, (\hat{p_1} - \hat{p_2}) + Z_{0.025}\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}} \right)$$

where n_1 is the size of the sample when the accrual rate of patients is exactly one patient per week, n_2 is the size of the sample when the accrual rate of patients is an average of one per week, $\hat{q_1} = 1 - \hat{p_1}$, and $\hat{q_2} = 1 - \hat{p_2}$.

Similarly, to test whether the difference between the means under the two accrual rates is greater than 4, we propose to test the hypotheses

$\mathrm{H}_0 : |\mu_1 - \mu_2| \leq 4$

$\mathrm{H}_1 : |\mu_1 - \mu_2| > 4$

where μ_1 is the mean when the arrival rate of patients is exactly one patient per week and μ_2 is the mean when the arrival rate of patients is an average of one per week.

In a manner similar to the comparison of p_1 and p_2, we could test the hypothesis at a level of 5% using a 95% CI for $\mu_1 - \mu_2$. Since both $n_1(= 5000)$ and $n_2(= 5000)$ are very large, we can assume that the sample means are Normally distributed. We do not know σ_1 and σ_2. However, both samples are so large that S_1^2 and S_2^2 should be very good estimates of σ_1^2 and σ_2^2 respectively.
The 95% CI estimate of the difference is:

$$\left((\bar{X}_1 - \bar{X}_2) - Z_{0.025}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + Z_{0.025}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

If the CI lies entirely outside the interval (-4, 4), we reject H_0. Otherwise, we retain H_0.

2.3.2 The Results of the Hypothesis Testing

As discussed earlier, the program by Ning (2009) gave similar results to the published results of HNL in some situations. Under Scenario 1 these were when n_0 equalled 1, for both the HNL and common designs, and for only the common design when n_0 equalled 30. Under Scenario 2, we matched the published results of HNL, in the HNL design when n_0 equalled 1. We therefore focused on the comparison of the results from the two recruitment regimes in these situations.

Table 2.5 shows an example of the results of the HNL approach with a recruitment rate of exactly one patient per week for Scenario 1 compared with the results of the modified approach with the accrual rate being an average of one patient per week. Using the results in Table 2.5, we carried out hypothesis testing as described in the previous subsection. The results showed that a 95% CI for the differences $p_1 - p_2$ was (-0.019, 0.003). It was clear that the CI lay completely within the interval (-0.05, 0.05). We therefore retained each H_0 and concluded that there was no significant difference between the probability of type I error obtained from the two arrival rates. Similarly, the results implied that 95%

CIs for the differences $\mu_1 - \mu_2$ for arm A and for arm B were (-0.7833, 1.2609) and (-0.8633, 1.1721) respectively. Hence they lay entirely inside the interval (-4, 4). We therefore retained each H_0 and concluded that the average numbers of patients in arm A (or arm B) did not depend on the two enrolment regimes.

Let us consider the average numbers of patients obtained under the two recruitment regimes. The results were such that a 95% CI for the differences $\mu_1 - \mu_2$ was $(-0.2623, 1.0487)$. Hence it lay entirely inside the interval (-4, 4). We therefore retained H_0 and concluded that the average numbers of patients in the trial were not based on the two enrolment regimes.

The results from the all the other situations mentioned in the first paragraph of this subsection follow similarly.

From the results mentioned above, we finally concluded that the differences between the statistical properties of the two recruitment regimes (i.e. exactly one new patient per week, and an average of one new patient per week) should not be considered to be practically significant. Therefore we will continue to follow the HNL practice of having exactly one arrival per week.

Table 2.5: Comparison of the results of the HNL approach with a recruitment rate of exactly one patient per week with the results of the modified approach with the accrual rate being an average of one patient per week.

Scenario	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	Source	HNL design	
											Average	Probability of selection as the superior arm
1	A	0.2	0.4	0.1	0.3	4	30	75	110	A	58.32	0.0408
										B	58.08	0.0432
	B	0.2	0.4	0.1	0.3	4	30	75	110	A	57.6034	0.0424
										B	57.45	0.0478
	total									A	115.93	
										B	115.53	

Let source A represent the results of the HNL approach with a recruitment rate of exactly one patient per week; and source B represent the results of the modified approach with the accrual rate for an average of one patient per week.

Chapter 3

Evaluation of Clinical Trial Designs

The aim of this chapter is to consider major criteria for evaluating a clinical trial design. This is because the evaluation of the design is an important objective, when designing a clinical trial. In this chapter, we consider two methods for evaluating and comparing clinical trial designs. They are consideration of the Operating Characteristic Curve and other design characteristics. The details of each method are described in the following sections.

3.1 Operating Characteristic curve

In this section, the Operating Characteristic curve will be considered to assess the performance of a clinical trial design.

Normally the Operating Characteristic (OC) curve is a useful tool in describing the capabilities of a sampling plan for discriminating between good or bad lots in quality control. The OC curve can be also applied to evaluate and compare adaptive designs. In this research we define the OC function to be the probability of accepting $H_0 : \mu_a - \mu_b = 0$ for a given value of the parameter(s), so it is equal to $1 - \alpha$ when H_0 is true, and 1 - power when H_0 is false. In a graph of an OC curve the horizontal axis shows the difference between the mean survival time of treatment A (μ_a) and the mean survival time of treatment B (μ_b), that is,

$\mu = \mu_a - \mu_b$. The vertical axis shows the probability of accepting H_0. Therefore this curve describes the discriminating ability of a design.

If Design A gives a steeper OC curve than a competing design (Design B), we will decide that Design A is the better design. That is, Design A has greater ability to distinguish an effective treatment from a less effective one. In this research, the hypotheses are

$H_0 : \mu_a - \mu_b = 0$

$H_1 : \mu_a - \mu_b \neq 0.$

The quantity α can be computed from the proportion of the time that we reject H_0 when H_0 is true, while power can be computed from the proportion of the time that we reject H_0 when H_0 is false. Initially we set the true value of $\mu = \mu_a - \mu_b$ equal to 0. This is the situation that H_0 is true, so we estimate α. Then we change the true value of μ. We use a wide domain for the true value of μ. In these situations where H_0 is false, we therefore calculate power.

3.2 Example of using the Operating Characteristic to compare designs

In this section, we provide an illustrative example of using the Operating Characteristic to compare designs.

We compare the Operating Characteristics (when H_0 is true) of a HNL design and an alternative design. In HNL, we had $\mu_x = \sum_{k=1}^{4} p_{x,k}\mu_{x,k}$ $(x = a, b)$. It is clear that there are many different choices of the $p_{x,k}$s and $\mu_{x,k}$s that give the same value of μ_x. Scenarios 1 (a) - 1 (d) of Table 3.1 give four examples for which $\mu_x = 53.3$. These scenarios, plus Scenarios 1 (e) - 1 (f) in Table 3.1, provide situations where $\mu = \mu_a - \mu_b = 0$. For these six scenarios, simulation was used to estimate the value of P(Retain $H_0|H_0$ true). We see that P(Retain $H_0|H_0$ true) was calculated from one minus (the probability that arm A was chosen plus the probability that arm B was chosen). For each simulation, data were generated for $N_{max} = 120$. In Scenarios 1 (a) - 1 (f), we performed hypothesis testing as described in Section 2.3.1 to test whether the two values of the OC were different. For example, Scenarios 1 (b) and 1 (e) both had $\mu_a - \mu_b = 0$, but the values of P(Retain $H_0|H_0$ true) differed by 0.1136. Recall that we decided that two values of the OC would be regarded as practically different if the absolute value of their difference exceeds or equals 0.05.

In Scenarios 1 (b) and 1 (e), the 95% CI for the differences $p_1 - p_2$ for arm A was $(0.1015, 0.1257)$. It can be seen that the CI for arm A did not lie within the interval $(-0.05, 0.05)$. It is also apparent from Table 3.1, that the value of P(Retain $H_0 : \mu_a = \mu_b|H_0$ true) varies considerably, both for the same and different values of $\mu_a - \mu_b$. Because of the variability in values of the OC when $\mu = 0$, we cannot regard P(Retain $H_0|H_0$ true) as being a function of $\mu_a - \mu_b$.

This suggests that the OC curve is not an appropriate method of comparing clinical trial designs.

Table 3.1: Comparison of the probabilities of retaining H_0 when H_0 is true for $N_{max} = 120$ under various scenarios

											HNL design		
Scenario	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	Average	Probability	Probability of retaining H_0
1 (a)	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	58.0964	0.0430	0.9172
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3	57.7900	0.0398	
	total										115.8864		
1 (b)	A	0.1	0.4	0.25	0.25	3	30	72	92	53.3	58.9752	0.0262	0.9460
	B	0.1	0.4	0.25	0.25	3	30	72	92	53.3	58.6004	0.0278	
	total										117.5756		
1 (c)	A	0.1	0.1	0.1	0.7	2	25	51	65	53.3	57.6624	0.0596	0.8886
	B	0.1	0.1	0.1	0.7	2	25	51	65	53.3	57.1330	0.0518	
	total										114.7954		
1 (d)	A	0.5	0.2	0.1	0.2	4	28	97	180	53.3	57.4834	0.0404	0.9202
	B	0.5	0.2	0.1	0.2	4	28	97	180	53.3	58.8514	0.0394	
	total										116.3348		
1 (e)	A	0.2	0.2	0.3	0.3	5	7	22	170	60	55.3086	0.0860	0.8324
	B	0.2	0.2	0.3	0.3	5	7	22	170	60	54.8254	0.0816	
	total										110.1340		
1 (f)	A	0.2	0.2	0.3	0.3	4	20	44	90	45	57.4108	0.0490	0.8962
	B	0.2	0.2	0.3	0.3	4	20	44	90	45	57.4350	0.0548	
	total										114.8458		

3.3 Design characteristics

In addition to the method in Section 3.1, we can evaluate clinical trial designs by considering various other design characteristics. In this section, several design characteristics are investigated and used as criteria to assess the clinical trial designs.

Jiang *et al.* (2013) proposed new designs for the phase II clinical trial. These designs use both a Bayesian decision-theoretic approach and a response-adaptive randomization procedure. They are used to evaluate the efficacy of two arms with a binary endpoint. The aim of Jiang *et al.* (2013) was to make an efficient and ethical design that can determine the efficacy of a treatment efficiently in order to screen out an inefficient treatment. In addition, for ethical reasons, this design should reduce the average sample number and increase the percentage of patients assigned to the superior treatment. The authors evaluated the performance of their design by considering several design characteristics - namely, the probability of Type I error, statistical power, the average number of patients and the average percentage of patients allocated to the superior treatment. In our research, we have the same goals. Consequently, in evaluating the procedure, we will use the same criteria as they did. In addition, comparing the percentage of patients assigned to the better treatment (PBA) will compare percentages, rather than actual counts, of patients. This is preferable if there are different total numbers of patients in the two trials.

Emerson *et al.* (2007) described the evaluation of a clinical trial design by considering design characteristics such as the probability of Type I error, and the power of the trial. They did it by giving an example. This example showed that the power of the trial increased as the treatment effect ($\mu = \mu_a - \mu_b$) increased.

Cheng and Shen (2005) considered an expected loss function, depending on the cost of each patient and the costs of making wrong decisions. They determined whether or not to terminate the trial by using this function as a criterion. They also used three design characteristics (the probability of Type I error, the power of the trial and the average sample size) to compare Bayesian adaptive designs and other group sequential designs.

Wunder *et al.* (2012) also compared adaptive designs by evaluating the design characteristics. In this paper, the authors considered not only the characteristics mentioned in Cheng and Shen (2005) but also the average number of deaths per trial and the proportion of early terminations.

Gaydos *et al.* (2009) stated that the assessment of the proposed adaptive designs could be performed by investigating design characteristics such as the Type I error rate, the average duration of the trial, the power of the trial and the average sample number.

In our research, one of the design characteristics (the probability of Type I error) is fixed. The probability of Type I error can be estimated from the proportion of the time that H_0 is rejected when H_0 is true. In this research, we use a probability of Type I error of approximately 0.05.

There are seven design characteristics that are employed as criteria to evaluate the clinical trial designs:

1. **The power of the test**

   The power of the test can be estimated from the proportion of the time that we reject H_0 when H_0 is false. A better design will give greater power.

2. **The percentage of patients assigned to the better treatment (PBA)**

   The percentage of patients assigned to the superior treatment can be estimated from the proportion of patients assigned to the superior treatment. We will decide that the design is more effective if it has a higher percentage of patients assigned to the superior treatment.

   We adopt this criterion from Jiang *et al.* (2013).

3. **The probability of early termination (PET)**

   The probability of early termination can be estimated from the proportion of times that the trial can be stopped before its scheduled finish. We will make a decision that a design is more effective if it has a higher probability of early termination.

4. **The average number of patients (ANP)**

   The average number of patients is the expected number of patients used in the trial. Under this criterion, if the design uses a smaller number of patients, it is more effective.

5. **The average number of deaths (AND)**

   The average number of deaths is the expected number of patients dying in the trial. Under this criterion, if the design gives a smaller numbers of deaths, it is more effective. Note that, in this research, the AND is the expected number of patients who occupy category 1. In addition, their survival times do not exceed the length of the trial.

6. **The average length of the trial (ALT)**

   The average length of the trial is the expected number of weeks that the trial runs. Under this criterion, a design that has a shorter length of trial will be a more effective design.

7. **Expectation cost**

   Expectation cost will be described in the next section.

For design characteristics 1 - 3, we decided that these design characteristics obtained from the two designs were different if the difference was greater than or equal to 0.05. In contrast, for design characteristics 4 - 6, we made a decision that these design characteristics obtained from the two designs were different if the difference was greater than or equal to 4.

3.3.1 Assumptions

- Of course either arm may be better, but for the case of simplicity it will be assumed that arm B is better if the two arms are not equally good.
- It is assumed that arm A is a standard treatment or a placebo. Arm B, however, is a new treatment or an experimental drug.

3.4 Expected cost

According to Emerson *et al.* (2011), the major aims of evaluating a clinical trial are to (1) obtain correct conclusions, (2) answer some ethical issues, (3) minimize the costs of the trials. By using design characteristics 1 and 2 as criteria for evaluating a design, the first aim can be achieved. We also achieve the second and third aims by using design characteristics 3 - 6 as criteria. Hence, we still require a criterion that can combine all objectives to assess a clinical trial simultaneously.

In order to achieve all purposes of evaluating a design simultaneously, in this section the expected cost of a design will be used as one of the design characteristics. We can evaluate and compare clinical trial designs by looking at the expected costs of the designs. This is because the factors that influence the expected costs are the probability of Type I error, the power of the trial and the average number of patients in the trial. The expected cost of a design combines the cost of patients, the cost of treatment, and the cost (e.g., the cost of lost opportunity) if we make a wrong decision.

Since the expected cost is based on the factors mentioned above, by using this criterion we can achieve all main aims of assessing a clinical design simultaneously.

In this research, two situations will be considered.

1. In situation 1, we have no prior knowledge of how arms A and B differ.
2. In situation 2, we suppose that treatment B is superior to treatment A.

In both situations, the hypotheses will be

$\mathrm{H}_0 : \mu_a \geq \mu_b$;

$\mathrm{H}_1 : \mu_a < \mu_b$.

We decided that a one-sided alternative hypothesis is more appropriate to our objective. This is because in Section 3.3.1, we assumed that arm A is a standard treatment or a placebo, whereas arm B is a new treatment or an experimental drug. Moreover, in situation 1, for a two sided hypothesis, if arm A is selected as a better treatment, it means that we make a wrong decision and we need to pay the cost for carrying out a Phase III trial. However, due to the assumption mentioned above, we do not need to perform the next phase if a standard treatment or a placebo is selected.

It should be noted here that from now on in this chapter we are only concerned about this one-sided alternative hypothesis. This is different from HNL. In HNL, for the first situation, a two-sided hypothesis was considered. In contrast, for the second situation, HNL considered this one-sided hypothesis.

Let us consider situation 1. If there is no difference between μ_a and μ_b, but we reject H_0, we make a wrong decision. We conclude that arm B is a better treatment, even though the efficacies of arm A and arm B are equal. As mentioned in Section 1.1, in general, a Phase II trial is used to compare a new treatment with a standard treatment or an experimental drug with a placebo. If a new treatment or an experimental drug (that is, arm B in this trial) is selected as a better treatment, we can proceed to a Phase III trial. This explains why the cost will be incurred if arm B is chosen as a better treatment and then a Phase III trial is undertaken.

Moving to situation 2, in this situation, it is supposed that treatment B is superior to treatment A. If we accept H_0, we make a wrong decision. We reject the superior treatment and incorrectly choose the wrong drug. Hence, the cost of lost opportunity will be incurred if arm B is not chosen as a better treatment. Let

- e_1 be the event that we reject H_0 when H_0 is true;
- e_2 be the event that we accept H_0 when H_0 is true;
- f_1 be the event that we accept H_0 when H_0 is false;
- f_2 be the event that we reject H_0 when H_0 is false;
- n_a be the average number of patients in arm A in this Phase II trial;
- n_b be the average number of patients in arm B in this Phase II trial;
- $n = n_a + n_b$ be the average number of patients in this Phase II trial;

For the cost, we adopt the notation from Cheng and Shen (2005), that is:

- K_0 = the cost of lost opportunity when we reject H_0 if H_0 is true;
- K_1 = the cost of lost opportunity when we accept H_0 if H_0 is false;
- K_2 = the cost of an individual patient excluding his/her cost of treatment.

We also introduce

- K_a = the cost of treatment A per patient;
- K_b = the cost of treatment B per patient.

The expected cost when H_0 is true can be defined as

$$E(\text{cost}) = \sum_{i=1}^{2} c(e_i)P(e_i), \tag{3.1}$$

where $c(e_1)$ is the associated cost of the trial when we reject H_0 if H_0 is true; $c(e_1) = nK_2 + n_aK_a + n_bK_b + K_0$;

$c(e_2)$ is the associated cost of the trial when we accept H_0 if H_0 is true; $c(e_2) = nK_2 + n_aK_a + n_bK_b$.

Therefore

$$\begin{aligned} E(\text{cost}) &= (nK_2 + n_aK_a + n_bK_b + K_0)P(e_1) \\ &\quad +(nK_2 + n_aK_a + n_bK_b)(1 - P(e_1)) \\ &= nK_2 + n_aK_a + n_bK_b + K_0P(e_1) \end{aligned}$$

From this formula, it can be clearly seen that the expected cost consists of two parts. Firstly, there is the constant cost that combines the patient's cost and the treatment cost. In addition, there is the cost for rejecting H_0 if H_0 is true, that is $K_0P(e_1)$.

The expected cost when H_0 is false can be expressed as

$$E(\text{cost}) = \sum_{i=1}^{2} c(f_i)P(f_i), \tag{3.2}$$

where $c(f_1)$ is the associated cost of the trial when we accept H_0 if H_0 is false and $c(f_1) = nK_2 + n_aK_a + n_bK_b + K_1$;

$c(f_2)$ is the associated cost of the trial when we reject H_0 if H_0 is false and $c(f_2) = nK_2 + n_aK_a + n_bK_b$.

Therefore

$$\begin{aligned} E(\text{cost}) &= (nK_2 + n_aK_a + n_bK_b + K_1)P(f_1) \\ &\quad +(nK_2 + n_aK_a + n_bK_b)(1 - Pr(f_1)) \\ &= nK_2 + n_aK_a + n_bK_b + K_1Pr(f_1) \end{aligned}$$

From the expected cost given above, we can see that again there are two main parts of the expected cost: the constant cost and the cost of lost opportunity from accepting H_0 if H_0 is false, that is $K_1Pr(f_1)$.

3.5 Example of using design characteristics to evaluate designs

In this section, an example of using this method to evaluate the design is shown. We examine the HNL design for six different treatment effects ($\mu = \mu_A - \mu_B$) under various values of N_{max}.

It should be noted that since we consider a one-sided alternative hypothesis (see Section 3.4), from now on in this chapter the probabilities of Type I error can be estimated from the proportion of the time that arm B is selected as a superior treatment in situation 1.

Table 3.2 displays Type I error rates for the HNL design under various values of N_{max}. In this research, we would like to obtain a Type I error rate of approximately 0.05. Unfortunately, by fixing the value of p_U, the probability of Type I error increased as N_{max} increased. Therefore, we performed many simulations to find a suitable cut-off p_U for each choice of N_{max}. The p_U was chosen if it gave a significance level of about 0.05. The results set out in Table 3.2 show that, as N_{max} increased, the ANP got larger. This is to be expected.

Tables 3.3 - 3.7 show the values of six design characteristics when H_0 is false for five different treatment effects under various values of N_{max} . The results displayed in Table 3.3 - 3.7 show that, for any given $p_{x,k}$s and $\mu_{x,k}$s, increasing N_{max} increased the ANP, the AND, the power of the trial, the PET, the PBA, and the ALT. As the N_{max} increased, the duration of the trials also increased. Consequently, the ALT and the ANP increased. When the ALT and the ANP increased, we got more information, so the trial can increase the probability of identification that arm B is the superior treatment. The power and the PBA increased. The ANP also led to larger AND and larger PBA because the AND was the proportion of the number of patients used. We can also stop trials early, when the ALT increased. Therefore, the PET increased.

In addition the power, the PBA and the PET increased, as the treatment effect increased. Moreover, the ANP, the ALT and the AND decreased, as the treatment effect increased. This is to be expected. If the treatment effect increases, the difference between two arms can be detected more easily and quickly. As a result, the power, the PBA and the PET increase. Due to an increase in the power, the ANP required to detect difference decreases. Additionally, we can more rapidly get enough evidence that arm B is better. Therefore the ALT decreases. As mentioned in design characteristic 4, in this research, the AND is the number of patients who fall in category 1 and do not survive more than the length of the trial. Since the ALT decreases at the end of the trial, more patients in this category are censored, and the AND becomes smaller.

Table 3.2: Comparison of the probabilities of Type I error for the HNL design, under various values of N_{max}

													HNL design		
Case	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	N_{max}	p_U	Average	Probability	Prob of Type I error
1 (a)	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	100	0.9700	48.0744	0.0476	0.0488
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3			47.9476	0.0488	
	total												96.0220		
1 (b)	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	120	0.9705	57.3314	0.0506	0.0512
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3			57.1970	0.0512	
	total												114.5284		
1 (c)	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	140	0.9727	67.3776	0.0540	0.0508
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3			66.1276	0.0508	
	total												133.9312		
1 (d)	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	160	0.9741	76.5222	0.0532	0.0574
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3			75.6804	0.0574	
	total												152.2026		
1 (e)	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	180	0.9760	85.5094	0.0532	0.0612
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3			84.7498	0.0612	
	total												170.2592		

Table 3.3: Comparison of the design characteristics when H_0 is false under various values of N_{max} when the treatment effect is 24.725

(a) Section 1

Parameters of the design												
Case	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	N_{max}	p_U
2(a)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.300	100	0.9700
	B	0.100	0.125	0.325	0.450	4	30	75	110	78.025		
	total											
2(b)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.300	120	0.9705
	B	0.100	0.125	0.325	0.450	4	30	75	110	78.025		
	total											
2(c)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.300	140	0.9727
	B	0.100	0.125	0.325	0.450	4	30	75	110	78.025		
	total											
2(d)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.300	160	0.9741
	B	0.100	0.125	0.325	0.450	4	30	75	110	78.025		
	total											
2(e)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.300	180	0.9760
	B	0.100	0.125	0.325	0.450	4	30	75	110	78.025		
	total											

(b) Section 2

		Properties of HNL design						
Case	Arm	ANP	Prob	AND	power	PET	PBA	ALT
2(a)	A	17.6092	0.0004	9.5254	0.4692	0.4318	0.7756	101.1838
	B	60.8466	0.4692					
	total	78.4558						
2(b)	A	19.4612	0.0006	10.8202	0.5126	0.4774	0.7855	111.6356
	B	71.2704	0.5126					
	total	90.7316						
2(c)	A	21.3498	0.0004	12.1848	0.5364	0.5070	0.7922	122.4484
	B	81.3786	0.5364					
	total	102.7284						
2(d)	A	22.8872	0.0004	13.4424	0.5598	0.5316	0.8006	133.4974
	B	91.8742	0.5598					
	total	114.7614						
2(e)	A	24.3630	0.0002	14.9844	0.5794	0.5518	0.8082	144.9730
	B	102.6820	0.5794					
	total	127.0450						

Let Prob be the probability of selecting arm A or B as the superior treatment.

Table 3.4: Comparison of the design characteristics when H_0 is false under various values of N_{max} when the treatment effect is 29.1

(a) Section 1

		Parameters of the design										
Case	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	N_{max}	P_U
3(a)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	100	0.9700
	B	0.100	0.125	0.200	0.575	4	30	75	110	82.40		
	total											
3(b)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	120	0.9705
	B	0.100	0.125	0.200	0.575	4	30	75	110	82.40		
	total											
3(c)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	140	0.9727
	B	0.100	0.125	0.200	0.575	4	30	75	110	82.40		
	total											
3(d)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	160	0.9741
	B	0.100	0.125	0.200	0.575	4	30	75	110	82.40		
	total											
3(e)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	180	0.9760
	B	0.100	0.125	0.200	0.575	4	30	75	110	82.40		
	total											

(b) Section 2

		Properties of HNL design						
Case	Arm	ANP	Prob	AND	power	PET	PBA	ALT
3(a)	A	15.2326	0.0010	8.6340	0.5602	0.5144	0.7934	93.1390
	B	58.4824	0.5602					
	total	73.7150						
3(b)	A	17.0928	0	9.7814	0.6038	0.5646	0.7964	101.3790
	B	66.8702	0.6038					
	total	83.9630						
3(c)	A	18.2242	0.0010	10.9084	0.6370	0.6040	0.8059	109.7084
	B	75.6442	0.6370					
	total	93.8684						
3(d)	A	19.7536	0.0002	12.1648	0.6464	0.6178	0.8127	120.7268
	B	85.6852	0.6464					
	total	105.4388						
3(e)	A	20.6170	0.0004	13.3430	0.6666	0.6428	0.8211	129.5314
	B	94.6264	0.6666					
	total	115.2434						

Let Prob be the probability of selecting arm A or B as the superior treatment.

Table 3.5: Comparison of the design characteristics when H_0 is false under various values of N_{max} when the treatment effect is 34.65

(a) Section 1

		Parameters of the design										
Case	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	N_{max}	P_U
4(a)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	100	0.9700
	B	0.050	0.100	0.250	0.600	4	30	75	110	87.95		
	total											
4(b)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	120	0.9705
	B	0.050	0.100	0.250	0.600	4	30	75	110	87.95		
	total											
4(c)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	140	0.9727
	B	0.050	0.100	0.250	0.600	4	30	75	110	87.95		
	total											
4(d)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	160	0.9741
	B	0.050	0.100	0.250	0.600	4	30	75	110	87.95		
	total											
4(e)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	180	0.9760
	B	0.050	0.100	0.250	0.600	4	30	75	110	87.95		
	total											

(b) Section 2

		Properties of HNL design						
Case	Arm	ANP	Prob	AND	power	PET	PBA	ALT
4(a)	A	11.9454	0.0002	4.9392	0.6582	0.6192	0.8208	81.9054
	B	54.7280	0.6582					
	total	66.6734						
4(b)	A	12.7060	0.0002	5.3568	0.7226	0.6832	0.8261	85.7410
	B	60.3630	0.7226					
	total	73.0690						
4(c)	A	13.6142	0	5.9392	0.7500	0.7152	0.8346	93.7046
	B	68.6984	0.7500					
	total	82.3126						
4(d)	A	14.4172	0	6.5216	0.7678	0.7404	0.8409	101.0028
	B	76.2016	0.7678					
	total	90.6188						
4(e)	A	15.0400	0	6.9746	0.7864	0.7636	0.8468	107.6224
	B	83.1264	0.7864					
	total	98.1664						

Let Prob be the probability of selecting arm A or B as the superior treatment.

Table 3.6: Comparison of the design characteristics of the HNL design when H_0 is false under various values of N_{max} when the treatment effect is 39.9

(a) Section 1

		Parameters of the design										
Case	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	N_{max}	P_U
5(a)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	100	0.9700
	B	0.050	0.100	0.100	0.750	4	30	75	110	93.20		
	total											
5(b)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	120	0.9705
	B	0.050	0.100	0.100	0.750	4	30	75	110	93.20		
	total											
5(c)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	140	0.9727
	B	0.050	0.100	0.100	0.750	4	30	75	110	93.20		
	total											
5(d)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	160	0.9741
	B	0.050	0.100	0.100	0.750	4	30	75	110	93.20		
	total											
5(e)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	180	0.9760
	B	0.050	0.100	0.100	0.750	4	30	75	110	93.20		
	total											

(b) Section 2

		Properties of HNL design						
Case	Arm	ANP	Prob	AND	power	PET	PBA	ALT
5(a)	A	10.2916	0.0002	4.3616	0.7490	0.7016	0.8293	72.2350
	B	50.0074	0.7490					
	total	60.2990						
5(b)	A	11.0252	0.0002	4.7306	0.803	0.7648	0.8345	76.0328
	B	55.5996	0.8030					
	total	66.6248						
5(c)	A	11.5088	0	5.1096	0.8184	0.7906	0.8430	81.6624
	B	61.7776	0.8184					
	total	73.2864						
5(d)	A	12.2880	0.0004	5.6092	0.8414	0.8106	0.8464	87.5680
	B	67.7040	0.8414					
	total	79.9920						
5(e)	A	12.9784	0	6.1374	0.8464	0.8202	0.8524	95.1284
	B	74.9580	0.8464					
	total	87.9364						

Let Prob be the probability of selecting arm A or B as the superior treatment.

Table 3.7: Comparison of the design characteristics of the HNL design when H_0 is false under various values of N_{max} when the treatment effect is 70.35

(a) Section 1

Parameters of the design												
Case	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	N_{max}	P_U
6(a)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	100	0.9700
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65		
	total											
6(b)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	120	0.9705
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65		
	total											
6(c)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	140	0.9727
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65		
	total											
6(d)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	160	0.9741
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65		
	total											
6(e)	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	180	0.9760
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65		
	total											

(b) Section 2

		Properties of HNL design						
Case	Arm	ANP	Prob	AND	power	PET	PBA	ALT
6(a)	A	8.6976	0	3.7558	0.9096	0.8634	0.8351	58.2240
	B	44.0624	0.9096					
	total	52.7600						
6(b)	A	8.9552	0	3.8530	0.9456	0.9168	0.8377	58.5000
	B	46.2168	0.9456					
	total	55.1720						
6(c)	A	8.9924	0	4.0296	0.9578	0.9428	0.8449	60.2696
	B	48.9892	0.9578					
	total	57.9816						
6(d)	A	9.5122	0	4.3556	0.9710	0.9564	0.8476	64.1436
	B	52.8874	0.9710					
	total	62.3996						
6(e)	A	9.5762	0	4.4882	0.9724	0.9636	0.8544	67.2426
	B	56.2104	0.9724					
	total	65.7866						

Let Prob be the probability of selecting arm A or B as the superior treatment.

3.5.1 The results of the expected cost

Refer to Section 3.4 for a description of the expected cost. From Table 3.2, in the event that H_0 is true, the expected cost (3.1) mentioned in Section 3.4 can be written as follows:

$N_{max} = 100$

$$E(\text{cost}) = 96.0220K_2 + 48.0744K_a + 47.9476K_b + 0.0488K_0,$$

$N_{max} = 120$

$$E(\text{cost}) = 114.5284K_2 + 57.3314K_a + 57.1970K_b + 0.0512K_0,$$

$N_{max} = 140$

$$E(\text{cost}) = 133.9312K_2 + 67.3776K_a + 66.1276K_b + 0.0508K_0,$$

$N_{max} = 160$

$$E(\text{cost}) = 152.9754K_2 + 76.5222K_a + 75.6804K_b + 0.0574K_0,$$

$N_{max} = 180$

$$E(\text{cost}) = 170.2592K_2 + 85.5094K_a + 84.7498K_b + 0.0612K_0.$$

For this example, it can be clearly observed that, when H_0 is true, for positive K_0, K_a, K_b and K_2, the expected cost increased significantly as N_{max} increased except when $N_{max} = 140$. As mentioned in Section 3.4, if H_0 is true, there are two principal parts of the expected cost: the constant cost and the cost from rejecting H_0 if H_0 is true. Again, when N_{max} increased, we observed a larger ANP. As a result, the constant cost is greater. In addition, the cost from rejecting H_0 if H_0 is true increases as N_{max} increases except when $N_{max} = 140$.

In other examples, where we obtain different coefficient of K_0, this might not be true.

Note that, in this trial, the probability of Type I error was obtained by carrying out the simulation. We chose the p_U if it gave a significance level of about 0.05. Hence, we might not obtain an upward trend of the expected cost as N_{max} increased. The Type I error rates for different $N_{max}s$ are not exactly the same. For a trial in which the Type I error rate is fixed, certainly, the expected cost increased significantly as N_{max} increased.

From Table 3.3, when the treatment effect is 24.725, in the event that H_0 is false, the expected cost (3.2) mentioned in Section 3.4 can be written as follows:
$N_{max} = 100$

$$E(\text{cost}) = 78.4558K_2 + 17.6092K_a + 60.8466K_b + 0.5308K_1,$$

$N_{max} = 120$

$$E(\text{cost}) = 90.7316K_2 + 19.4612K_a + 71.2704K_b + 0.4874K_1,$$

$N_{max} = 140$

$$E(\text{cost}) = 102.7284K_2 + 21.3498K_a + 81.3786K_b + 0.4636K_1,$$

$N_{max} = 160$

$$E(\text{cost}) = 114.7614K_2 + 22.8872K_a + 91.8742K_b + 0.4402K_1,$$

$N_{max} = 180$

$$E(\text{cost}) = 127.0450K_2 + 24.3630K_a + 102.6820K_b + 0.4206K_1,$$

From Table 3.4, when the treatment effect is 29.1, in the event that H_0 is false, the expected cost (3.2) mentioned in Section 3.4 can be written as follows:

$N_{max} = 100$

$$E(\text{cost}) = 73.7150K_2 + 15.2326K_a + 58.4824K_b + 0.4398K_1,$$

$N_{max} = 120$

$$E(\text{cost}) = 83.9630K_2 + 17.0928K_a + 66.8702K_b + 0.3962K_1,$$

$N_{max} = 140$

$$E(\text{cost}) = 93.8684K_2 + 18.2242K_a + 75.6442K_b + 0.0.363K_1,$$

$N_{max} = 160$

$$E(\text{cost}) = 105.4388K_2 + 19.7536K_a + 85.6852K_b + 0.3536K_1,$$

$N_{max} = 180$

$$E(\text{cost}) = 115.2434K_2 + 20.6170K_a + 94.6264K_b + 0.3334K_1.$$

The equations from the other Tables follow similarly.

It can be clearly observed from the results mentioned above that, if H_0 is false, the coefficients of all costs except K_1 increased as N_{max} increased. However the coefficient of K_1 decreased as N_{max} increased. When the sample size increases, we obtain more information. As a result, there is an increase in the ability to detect the differences between the two arms. The obtained power of the test will increase. Therefore, the coefficient of K_1, that is, (1- power), decreases.

As mentioned in Section 3.4, if H_0 is false, the expected cost consists of two main parts: a constant cost and the cost of lost opportunity from accepting H_0 if H_0 is false. As N_{max} increases, we can reduce the cost of lost opportunity $K_1 Pr(f_1)$ from accepting H_0 if H_0 is false; however, the constant cost $(nK_2 +$

$n_aK_a + n_bK_b$) increases.

In practice, we need to determine the values of K_1, K_2, K_a and K_b. If the value of K_1 is considerably larger than the values of K_2, K_a and K_b, when the N_{max} is large, the cost of lost opportunity can be reduced by more than the increase in constant cost. Consequently, in this situation, the expected cost decreases as N_{max} increases.

For illustrative purposes, a simple example is given here. Suppose that the values of K_2, K_a and K_b are 1 unit, whereas the value of K_1 is 2000 units. It can be illustrated by the expected cost from Table 3.3, for N_{max} of 100 and 180.

$N_{max} = 100$

$$\begin{aligned} E(\text{cost}) &= 73.7150(1) + 15.2326(1) + 58.4824(1) + 0.4398(2000), \\ &= 1027.03. \end{aligned}$$

$N_{max} = 180$

$$\begin{aligned} E(\text{cost}) &= 115.2434(1) + 20.6170(1) + 94.6264(1) + 0.3334(2000), \\ &= 897.29. \end{aligned}$$

In this case, for an N_{max} of 100, the expected cost is 1027.03 units, but the expected cost is 897.29 units when N_{max} is 180. Therefore, in this case, we can reduce the expected cost slightly when we use a larger N_{max}.

However, if the value of K_1 is not substantially larger than the values of K_2, K_a and K_b, the expected cost increases as N_{max} increases because only the coefficient of K_1 decreases as N_{max} increases. In addition, the coefficient of K_1, which is less than 1, is considerably lower than the coefficients of K_2, K_a and K_b. In this situation, if we increase N_{max}, the power will increase so the cost of lost opportunity can be reduced slightly. However, we need to spend a lot of money to meet the costs of patients and treatments.

Furthermore the expected cost decreases as the treatment effect increases. If the treatment effect increases, we can detect the difference between the efficacies of the two treatments more easily and quickly. The power of the test increases. In addition, a smaller sample size is required.

Although increasing N_{max} can increase the power, there are some limitations that arise from using a large N_{max}. Firstly, it may be difficult to find enough patients to participate in the trial, such as in a rare disease. In other situations, we have adequate numbers of patients; however, there may be limits on the facilities available to treat them. For example, if the trials require a specific tool, this tool may not be sufficiently available to treat the patients in a large trial. When planning a clinical trial, we must determine the precision that we require and our resources such as budget, patients, tools and time, because larger sample sizes need more resources.

3.6 Example of using design characteristics to compare clinical trial designs

In this section, we provide simulation results to show how to use design characteristics to compare three designs: the HNL design, the common design and an equal randomization (ER) design. All designs were produced by using Ning (2009). Table 3.8 shows some features of the HNL design, the common design for $n_0 = 1$ and the ER design.

In order to obtain an ER design, we did simulation by using Ning (2009) where n_0 is equal to N_{max}. Therefore, the Dir $(\gamma_{x,1}+n_{x,1}, \gamma_{x,2}+n_{x,2}, \gamma_{x,3}+n_{x,3}, \gamma_{x,4}+n_{x,4})$ and the IG$(\alpha'_{x,k} = \alpha_{x,k} + \sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}, \beta'_{x,k} = \beta_{x,k} + \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})$ distributions were calculated after all patients had been recruited to the trial. Then we simulated the $(p_{x,1}, ..., p_{x,4})$ from the first distribution and the $(\mu_{x,1}, ..., \mu_{x,4})$ from the second. After that, the program computed $\mu_x = \sum_{k=1}^{4} p_{x,k}\mu_{x,k}$ and evaluated $p = Pr(\mu_a >$

Table 3.8: Comparison of some features of the HNL design, the common design for $n_0 = 1$ and the ER design

Features	HNL design	Common design	ER design
What is the method used to assign treatment to patients?	Response-Adaptive Randomisation	Response-Adaptive Randomisation	Equal randomization
Was there an interim analysis of the trial? How often?	Yes, every week during the enrolment period.	Yes, every week during the enrolment period.	No, the analysis was performed at the end of the trial.
Can the trial be stopped early?	Yes	Yes	No

$\mu_b \mid data$). Finally treatment A (or B) was chosen as the better treatment if $p_A > p_U$ (or $p_A < p_L$). A property of this design is that this trial cannot be stopped early.

This design is a Bayesian design even though it is not adaptive because it uses a Bayesian method in the treatment decision process as mentioned above.

Table 3.9 illustrates the Type I error rate for the HNL design, the ER design and the common design when N_{max} is 120. The design characteristics when H_0 is false are shown in Table 3.10 for the HNL design, the ER design and the common design when the treatment effect is 70.35 and N_{max} is 120. We note that, for the common design, the AND cannot be calculated because in this design, the short-term response was not classified into four categories, so we cannot say which patients died. It should be also noted that in this section, $n_0 = 1$ for the HNL and common designs. In contrast, $n_0 = 120$, for the ER design.

The results set out in Table 3.9 show that, in order to obtain a Type I error rate of approximately 0.05, the ER design required the lowest value of p_U compared to the other two designs. The common design needed the highest value of p_U.

The results shown in Table 3.10 illustrate that the HNL design required the lowest ANP and the shortest ALT. It also gave the highest PET. Moreover, the AND obtained from this design is less than one obtained from the ER design. In contrast, the ER design had the highest ANP and the longest ALT. Furthermore, it gave the highest power. In this design, the PET is zero because the design cannot be terminated early. As expected, we got the lowest power from the common design.

Similar simulations were done for other set of parameter which gave the same general conclusions.

One of the techniques used in the HNL design is an interim analysis. As mentioned in Section 1.3, an interim analysis offers opportunities for early termination of the trial. In the HNL design, the trial can stop early if one arm is demonstrably better than another arm. This can reduce the ANP and the ALT. Since the ALT decreases, the AND becomes small. The information about a short term response can help us to get sufficient evidence speedily. Therefore, the PET for the HNL design is highest compared to the other two designs.

In this research, all designs are Bayesian designs. The posterior probabilities are used to determine the superior treatment. We see that these probabilities are based on all available information obtained from the trial. Since the ER design has the highest ANP and the longest ALT, it obtains more information than the other two designs. The greater information will help it to make better decisions. This might explain why we get the highest power from the ER design.

From the results in Table 3.10, we carried out hypothesis testing as described in Section 2.3.1, to test whether these design characteristics for the HNL design, the ER design and the common design were different.

Let

p_{er} denote the probability (e.g. the power) obtained from the ER design;

p_h denote the corresponding probability obtained from the HNL design;

p_c denote the corresponding probability obtained from the common design;

μ_{er} denote the mean (e.g the ANP) obtained from the ER design;

μ_h denote the corresponding mean obtained from the HNL design;

μ_c denote the corresponding mean obtained from the common design;

The results of hypothesis testing (all based on 5,000 simulations) when comparing the design characteristics obtained from the HNL design and the ER design are as follows:

- 95% CIs for the differences $p_{er} - p_h$ for the power, for the PET and for the PBA were $(0.0255, 0.0405)$, $(-0.9245, -0.9091)$ and $(-0.3549, -0.3205)$ respectively. It can be seen that the CI for the the difference in power lay completely inside the interval $(-0.05, 0.05)$. The CIs for the differences in PET and the PBA, on the other hand, lay entirely outside the interval $(-0.05, 0.05)$. Consequently we concluded that there was no difference between the powers, whether or not these powers are associated with the HNL design or the ER design. However, the PETs and the PBAs obtained from the two designs were different.

- Similarly, 95% CIs for the differences $\mu_{er} - \mu_h$ for the ANP, for the AND and for the ALT were $(63.9124, 65.7436)$, $(10.9697, 11.2330)$ and $(100.3759, 102.6241)$ respectively. Hence they lay completely outside the interval (-4, 4). We concluded that the ANPs, the ANDs and the ALTs obtained from the two designs were different.

Consider the results of hypothesis testing when comparing the design characteristics obtained from the HNL design and the common design. Also consider the results of hypothesis testing when comparing the design characteristics obtained from the ER design and for the common design. We found that all 95% CIs for $p_h - p_c$ (not shown here) and $p_{er} - p_c$ (not shown here) lay entirely outside the interval $(-0.05, 0.05)$. Also, all CIs for $\mu_h - \mu_c$ (not shown here) and $\mu_{er} - \mu_c$ (not shown here) lay completely outside the interval (-4, 4). It can therefore be concluded that the design characteristics obtained from the HNL and the common designs were different. Similarly, we conclude that the design characteristics obtained from the the ER and the common designs were different.

Further consideration occurs in Section 3.8.1

Table 3.9: Comparison of the probabilities of Type I error for the HNL design, the ER and the common design when N_{max} is 120

Source	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	p_U	Average	Probability	Prob of Type I error
A	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	0.9705	57.3314	0.0506	0.0512
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3		57.1970	0.0512	
	total											114.5284		
B	A	0.2	0.4	0.1	0.3	4	30	75	110	53.3	0.9050	60.0446	0.0552	0.0474
	B	0.2	0.4	0.1	0.3	4	30	75	110	53.3		59.9554	0.0474	
	total											120		
C	A	1				53.3				53.3	0.993	57.3608	0.0444	0.0482
	B	1				53.3				53.3		57.6874	0.0482	
	total											115.0482		

Let source A represent the results of the HNL design, source B represent the results of the ER design and source C represent the results of the common design.

Table 3.10: Comparison of the design characteristics when H_0 is false for the HNL design, the ER design and the common design when the treatment effect is 70.35 and N_{max} is 120

(a) Section 1

Parameters of the design											
Source	Arm	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	μ_x	P_U
A	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	0.9705
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65	
total											
B	A	0.200	0.400	0.100	0.300	4	30	75	110	53.30	0.9050
	B	0.050	0.100	0.100	0.750	3	20	90	150	123.65	
total											
B	A	1				53.30				53.30	0.993
	B	1				123.65				123.65	
total											

(b) Section 2

		Properties of HNL design						
Source	Arm	ANP	Prob	AND	power	PET	PBA	ALT
A	A	8.9552	0	3.8530	0.9456	0.9168	0.8377	58.5000
	B	46.2168	0.9456					
	total	55.1720						
B	A	59.9992	0	14.9544	0.9786	0	0.5000	160
	B	60.0008	0.9786					
	total	120						
C	A	20.491	0.0012	NA	0.8574	0.7666	0.7517	128.4708
	B	62.0326	0.8574					
	total	82.5236						

Let source A be the results of the HNL design, source B be the results of the ER design, source C be the results of the common design, N/A be Not Available and Prob be the probability of selecting arm A or B as the superior treatment..

3.6.1 The results of expected cost

Refer to Section 3.4 for a description of Expected cost. From Table 3.9, in the event that H_0 is true, the expected cost (3.1) mentioned in Section 3.4 can be written as follows:

The HNL design

$$E(\text{cost}) = 114.5284K_2 + 57.3314K_a + 57.1970K_b + 0.0512K_0,$$

The ER design

$$E(\text{cost}) = 120K_2 + 60.0446K_a + 59.9554K_b + 0.0474K_0,$$

The common design

$$E(\text{cost}) = 115.0482K_2 + 57.3608K_a + 57.6874K_b + 0.0482K_0,$$

As mentioned in Section 3.4, if H_0 is true, there are two principal parts of the expected cost: the constant cost and the cost of lost opportunity from rejecting H_0 if H_0 is true. Using the HNL and the common designs can reduce the constant cost compared to the ER designs. For this example, however, they have a higher cost from rejecting H_0 if H_0 is true than the ER design. If the value of K_0 is considerably larger than the values of K_2, K_a and K_b, using the ER design has a lower cost than using the other two designs. Otherwise, the HNL and the common designs have lower cost compared to the ER design.

From Table 3.9, when treatment effect = 70.35, in the event that H_0 is false, the expected cost (3.2) mentioned in Section 3.4 can be written as follows:

The HNL design

$$E(\text{cost}) = 55.1720K_2 + 8.9552K_a + 46.2168K_b + 0.0544K_1,$$

The ER design

$$E(\text{cost}) = 120.0000K_2 + 59.9992K_a + 60.0008K_b + 0.0214K_1,$$

The common design

$$E(\text{cost}) = 82.5236K_2 + 20.491K_a + 62.0326K_b + 0.14764K_1,$$

As mentioned in Section 3.4, if H_0 is false, there are two principal parts of the expected cost: the constant cost and the cost of lost opportunity from accepting H_0 if H_0 is false. Using the HNL design can reduce the constant cost substantially compared to the other two designs. On the other hand, it has a higher cost of lost opportunity from accepting H_0 if H_0 is false than the ER design. If the value of K_1 is significantly larger than the values of K_2, K_a and K_b, using the ER design has a lower cost than using the HNL design. Otherwise, the HNL design requires lowest cost compared to the other two designs.

3.7 The expected total costs of the designs

In Section 3.4, the expected cost is one of the design characteristics suggested to evaluate the performance of a design. In the previous section, we determined the expected cost when H_0 is true, and the expected cost when H_0 is false, separately. In real life, however, this is hardly possible. As a result, in this section we suggest an alternative criterion by examining the expected cost in both situations simultaneously.

Parnell (2002) proposed a new criterion for evaluating experimental designs. In his research, the expected total costs of the trials were investigated. These were formulated by considering the hypotheses

$\mathrm{H}_0 : \mu_a = \mu_b;$

$\mathrm{H}_1 : \mu_a \neq \mu_b.$

According to Parnell (2002, p. 28), the expected total cost of a design can be defined as

$$E(\text{total cost}) = C_E + E(C_0), \tag{3.3}$$

where C_E is the direct cost of conducting the trials, and C_0 is the cost that results from the outcomes of the clinical trials. Hence $E(C_0)$ is the weighted sum of all possible costs associated with the four possible outcomes, given by

$$E(C_0) = P_{H_0} \times [\alpha \times C_1 + (1-\alpha) \times C_2] + (1 - P_{H_0}) \times [\beta \times C_3 + (1-\beta) \times C_4], \tag{3.4}$$

where P_{H_0} is the probability that H_0 is true;

α is the probability of rejecting H_0 when H_0 is true;

β is the probability of accepting H_0 when H_0 is false;

C_1 is the associated cost of the trial when we reject H_0 if H_0 is true;

C_2 is the associated cost of the trial when we accept H_0 if H_0 is true;

C_3 is the associated cost of the trial when we accept H_0 if H_0 is false;

C_4 is the associated cost of the trial when we reject H_0 if H_0 is false.

In this section, we use the same notation as in Section 3.4, namely,

K_0 = the cost when we reject H_0 if H_0 is true;

K_1 = the cost of lost opportunity when we accept H_0 if H_0 is false;

K_2 = the cost of an individual patient, excluding his/her cost of treatment;

K_a = the cost of treatment A per patient;

K_b = the cost of treatment B per patient.

As in Section 3.4, the determination of costs of the designs is considered for two situations.

- In situation 1, there is no prior knowledge of how arms A and B differ.
- In situation 2, treatment B is supposed to be superior to treatment A.

Again, in these situations, the hypotheses will be

$\mathrm{H}_0 : \mu_a \geq \mu_b$;

$\mathrm{H}_1 : \mu_a < \mu_b$.

In this research, we use an adaptive design. Our design also can terminate early if we have enough evidence that one arm is better. In the two situations, the numbers of patients required are different. For example, in situation 2, since we suppose that treatment B is superior to treatment A, the trial may stop earlier than in situation 1 where the efficacies of arm A and arm B are equal. Consequently, in situation 2, the trial may require smaller number of patients than in situation 1.

In this section, we introduce the notation

n_{a1} = the average number of patients in arm A in this Phase II trial in situation 1;

n_{a2} = the average number of patients in arm A in this Phase II trial in situation

2;

$n_a = P_{H_0} \times n_{a1} + (1 - P_{H_0}) \times n_{a2}$ = the expected number of patients in arm A in this Phase II trial;

n_{b1} = the average number of patients in arm B in this Phase II trial in situation 1;

n_{b2} = the average number of patients in arm B in this Phase II trial in situation 2;

$n_b = P_{H_0} \times n_{b1} + (1 - P_{H_0}) \times n_{b2}$ = the expected number of patients in arm B in this Phase II trial.

$n_1 = n_{a1} + n_{b1}$ = the average number of patients in this Phase II trial in situation 1;

$n_2 = n_{a2} + n_{b2}$ = the average number of patients in this Phase II trial in situation 2;

$n = n_a + n_b$ = the average number of patients in this Phase II trial.

Note that n can be computed by using the formula above or by using $n = P_{H_0} \times n_1 + (1 - P_{H_0}) \times n_2$. The two different ways give exactly the same answer.

We adopt formula (3.3) from Parnell (2002). By using the definition of Parnell, C_E, C_1, C_2, C_3, C_4 are given as

$$\begin{aligned} C_E &= nK_2 + n_aK_a + n_bK_b; \qquad (3.5) \\ C_1 &= n_1K_2 + K_0; \\ C_2 &= n_1K_2; \\ C_3 &= n_2K_2 + K_1; \\ C_4 &= n_2K_2. \end{aligned}$$

Hence $C_{E1} = n_1K_2 + n_{a1}K_a + n_{b1}K_b$ = the direct cost of conducting the trials in situation 1;

$C_{E2} = n_2K_2 + n_{a2}K_a + n_{b2}K_b$ = the direct cost of conducting the trials in situation

2.

Note that by replacing the n, n_a, n_b, (3.5) can be rewritten as

$$\begin{aligned}
C_E &= nK_2 + n_aK_a + n_bK_b \\
&= K_2 \times [P_{H_0} \times n_1 + (1 - P_{H_0}) \times n_2] + K_a \times [P_{H_0} \times n_{a1} + (1 - P_{H_0}) \times n_{a2}] \\
&\quad +K_b \times [P_{H_0} \times n_{b1} + (1 - P_{H_0}) \times n_{b2}] \\
&= P_{H_0} \times [n_1K_2 + n_{a1}K_a + n_{b2}K_b] + (1 - P_{H_0})[n_2K_2 + n_{a2}K_a + n_{b2}K_b] \\
&= P_{H_0} \times C_{E1} + (1 - P_{H_0}) \times C_{E2}
\end{aligned}$$

The associated costs of the trial for these four outcomes are displayed in Table 3.11.

Table 3.11: The associated costs of the trial for four outcomes as given by using the definition of Parnell (2002).

Situation	Outcome	
	Accept H_0	Reject H_0
H_0 is true	$C_2 = n_1K_2$	$C_1 = n_1K_2 + K_0$
H_0 is false	$C_3 = n_2K_2 + K_1$	$C_4 = n_2K_2$

In this research, we focus on a trial which is carried out to compare two treatments (arms A and B). Consequently, three outcomes can occur. Firstly, arm A is chosen as the superior treatment. Secondly, arm B is chosen as the superior treatment. Thirdly, the trial does not reach a conclusion, if no arm is selected as a better treatment during the trial. Since there are two situations and three different outcomes in each situation, in this research, six possible outcomes will exist instead of just four outcomes as in Parnell (2002).

As mentioned in Section 1.1, in a Phase II trial, an experimental treatment or a new treatment is compared with a standard treatment or a placebo. If a new treatment or an experimental drug is chosen as a better treatment, we can proceed to a Phase III trial.

Recall that it is assumed that arm A is a standard treatment or a placebo, whereas arm B is a new treatment or an experimental drug.

Let us consider situation 1. If the efficacies of arm A and arm B are equal, when we reject H_0, we make a wrong decision and decide that one arm is superior to another arm. If arm B is selected as a superior treatment, then we carry out the next phase. In reality, when both arms have the same efficacies, the existing drug should still be used and consequently, the next phase should not be undertaken. In this case, we pay the cost of conducting the Phase III trial unnecessarily.

Now consider situation 2. If treatment B is superior to treatment A, and we accept H_0, then the truly superior treatment is not chosen. Treatment A might be selected, or else the trial might not reach a conclusion. Hence, we incur the cost of lost opportunity for rejecting the superior treatment and incorrectly choosing the wrong drug.

From now on let K_0 = the cost when arm B is selected as a better treatment in situation 1; K_1 = the cost of lost opportunity when arm A is selected as a better treatment or a trial does not reach a conclusion in situation 2.

The cost of the trial for these six outcomes is given in Table 3.12.

Table 3.12: The associated cost of the trial for the six outcomes as given in this research.

Situation	Outcome		
	Arm A is selected as the better treatment	Arm B is selected as the better treatment	Conclusion not reached
H_0 is true	n_1K_2	$n_1K_2 + K_0$	n_1K_2
H_0 is false	$n_2K_2 + K_1$	n_2K_2	$n_2K_2 + K_1$

It can be seen from Table 3.12 that in both situations, the cost of the trial is the same, if either arm A is selected as the better treatment or else the trial does not reach a conclusion, .

For this reason, only the associated costs of the trial for four possible outcomes will be considered.

Therefore,

$E(C_0)$ is the weighted sum of all possible costs of the four possible outcomes.

$E(C_0)$ was given by

$$E(C_0) = P_{H_0} \times [\alpha \times C_1 + (1-\alpha) \times C_2] + (1 - P_{H_0}) \times [\beta \times C_3 + (1-\beta) \times C_4] \quad (3.6)$$

where n_a, n_b, n, n_1, n_2 have the same meaning as in (3.5);

α is the probability of selecting arm B as a better treatment, when really, the efficacies of arm A and arm B are equal;

β is the probability of selecting arm A as a better treatment or having an inconclusive trial, when arm B is truly the better arm;

C_1 is the associated cost of the trial when arm B is selected as a better treatment, but in reality, the efficacies of arm A and arm B are equal:

$$C_1 = n_1 K_2 + K_0;$$

C_2 is the associated cost of the trial when arm A is selected as a better treatment or a trial does not reach a conclusion, but in reality, the efficacies of arm A and arm B are equal:

$$C_2 = n_1 K_2;$$

C_3 is the associated cost of the trial when arm A is selected as a better treatment or a trial does not reach a conclusion, but arm B is really the better arm:

$$C_3 = n_2 K_2 + K_1;$$

C_4 is the associated cost of the trial when arm B is selected as a better treatment and arm B is truly the better arm:

$$C_4 = n_2 K_2;$$

The associated costs of the trial for these four outcomes are shown in Table 3.13.

Table 3.13: The associated costs of the trial for four outcomes.

Situation	Outcome	
	Arm A is selected as a better treatment or Conclusion not reached	Arm B is selected as a better treatment
H_0 is true	$C_2 = n_1K_2$	$C_1 = n_1K_2 + K_0$
H_0 is false	$C_3 = n_2K_2 + K_1$	$C_4 = n_2K_2$

Therefore

$$\begin{aligned} E(C_0) &= P_{H_0} \times [\alpha \times (n_1K_2 + K_0) + (1-\alpha) \times (n_1K_2)] \\ &\quad +(1-P_{H_0}) \times [\beta \times (n_2K_2 + K_1) + (1-\beta) \times (n_2K_2)], \\ &= P_{H_0} \times (n_1K_2 + \alpha K_0) + (1-P_{H_0}) \times (n_2K_2 + \beta K_1) \\ &= K_2 \times [P_{H_0} \times n_1 + (1-P_{H_0})n_2] + P_{H_0}\alpha K_0 + (1-P_{H_0})\beta K_1 \\ &= nK_2 + P_{H_0}\alpha K_0 + (1-P_{H_0})\beta K_1 \qquad (3.7) \end{aligned}$$

$E(C_0)$ consists of three main parts. Firstly, the cost of all patients (excluding treatment costs). Secondly, there is the weighted cost from rejecting H_0 if H_0 is true (in situation 1), that is $P_{H_0}\alpha K_0$. Moreover, there is the weighted cost of lost opportunity from accepting H_0 if H_0 is false (in situation 2), that is $(1-P_{H_0})\beta K_1$.

Then (3.3) can be rewritten in the form

$$E(\text{total cost}) = 2nK_2 + n_aK_a + n_bK_b + P_{H_0}\alpha K_0 + (1-P_{H_0})\beta K_1 \qquad (3.8)$$

3.8 Example of using the expected total cost

In this section, our main concern is to compare the expected total costs of the HNL and ER designs. This is because in Section 3.6, we found that these designs gave similar power whereas the common designs gave significantly lower power. This means that the HNL and ER designs are competitive.

Since the formulae for these expected total costs can be written as linear expressions, we can use algebra to identify the situation where one design is optimal or the situation where both designs are equally optimal.

Suppose that K_0, K_1, K_2, K_a, K_b have the same values for the two designs.

From equation (3.8), the expected total cost of the HNL design is

$$\begin{aligned} E(\text{total cost}) &= 2n_{HNL}K_2 + n_{aHNL}K_a + n_{bHNL}K_b + P_{H_0}\alpha_{HNL}K_0 \\ &+ (1 - P_{H_0})\beta_{HNL}K_1 \end{aligned} \tag{3.9}$$

(by using the formulae at page 77)

$$\begin{aligned} &= 2n_{2HNL}K_2 + n_{a2HNL}K_a + n_{b2HNL}K_b + \beta_{HNL}K_1 \\ &+ P_{H_0} \times [2K_2(n_{1HNL} - n_{2HNL}) + \alpha_{HNL}K_0 - \beta_{HNL}K_1 \\ &+ K_a(n_{a1HNL} - n_{a2HNL}) + K_b(n_{b1HNL} - n_{b2HNL})] \\ &= A + BP_{H_0}; \end{aligned} \tag{3.10}$$

where $A = 2n_{2HNL}K_2 + n_{a2HNL}K_a + n_{b2HNL}K_b + \beta_{HNL}K_1$ and $B = 2K_2(n_{1HNL} - n_{2HNL}) + \alpha_{HNL}K_0 - \beta_{HNL}K_1 + K_a(n_{a1HNL} - n_{a2HNL}) + K_b(n_{b1HNL} - n_{b2HNL})$.

Similarly, the expected total cost of the ER design is

$$\begin{aligned} E(\text{total cost}) &= 2n_{ER}K_2 + n_{aER}K_a + n_{bER}K_b + P_{H_0}\alpha_{ER}K_0 \\ &+ (1 - P_{H_0})\beta_{ER}K_1 \end{aligned} \tag{3.11}$$

(again by using the formulae at page 77)

$$
\begin{aligned}
&= 2n_{2ER}K_2 + n_{a2ER}K_a + n_{b2ER}K_b + \beta_{ER}K_1 \\
&+ P_{H_0} \times [2K_2(n_{1ER} - n_{2ER}) + \alpha_{ER}K_0 - \beta_{ER}K_1 \\
&+ K_a(n_{a1ER} - n_{a2ER}) + K_b(n_{b1ER} - n_{b2ER})] \\
&= C + DP_{H_0}; \qquad (3.12)
\end{aligned}
$$

where $C = 2n_{2ER}K_2 + n_{a2ER}K_a + n_{b2ER}K_b + \beta_{ER}K_1$ and $D = 2K_2(n_{1ER} - n_{2ER}) + \alpha_{ER}K_0 - \beta_{ER}K_1 + K_a(n_{a1ER} - n_{a2ER}) + K_b(n_{b1ER} - n_{b2ER})$.

Suppose that $y = E(\text{total cost})$ and $x = P_{H_0}$.

We have $y_{HNL} = A + Bx$ and $y_{ER} = C + Dx$.

There are two main kinds of possibilities that will be considered.

1. If $B = D$, the two lines are parallel. In this case $y_{HNL} > y_{ER}$ if $A > C$ and $y_{HNL} < y_{ER}$ if $A < C$. Furthermore, these lines will be the same if $A = C$.

2. If $B \neq D$, the two lines intersect when $A + Bx = C + Dx$. The point of intersection is at $x = \frac{C-A}{B-D}$.

   For $B \neq D$, let us consider the two cases:

   (a) $\frac{C-A}{B-D}$ lies between 0 and 1, and

   (b) $\frac{C-A}{B-D}$ does not lie between 0 and 1.

For the case that $\frac{C-A}{B-D}$ lies between 0 and 1, we consider

1. If $A = C$, the two lines will intersect at $x = 0$ and thereafter (for $x > 0$) $y_{HNL} > y_{ER}$ if $B > D$, or $y_{HNL} < y_{ER}$ if $B < D$,

2. If $A \neq C$, then there are two cases to consider.

   (a) For $A > C$, $y_{HNL} > y_{ER}$ for $x < \frac{C-A}{B-D}$ and $y_{HNL} < y_{ER}$ for $x > \frac{C-A}{B-D}$.

   (b) For $A < C$, $y_{HNL} < y_{ER}$ for $x < \frac{C-A}{B-D}$ and $y_{HNL} > y_{ER}$ for $x > \frac{C-A}{B-D}$

For the case where $\frac{C-A}{B-D}$ does not lie between 0 and 1, one method has greater y than another method in all practical circumstances.

As a simple example, suppose $K_0 = 490,000, K_1 = 440,000$, $K_2 = 50$ and $K_a = K_b = 100$.

From the results in Table 3.9 - 3.10 in Section 3.6, we begin by calculating A, B, C and D.

$$\begin{aligned}
A &= (2 \times 55.17 \times 50) + (8.96 \times 100) + (46.22 \times 100) + (0.05 \times 440,000) \\
&= 33,035 \\
B &= 100 \times (114.52 - 55.17) + (0.05 \times 490,000) - (0.05 \times 440,000) \\
&\quad +100 \times (57.33 - 8.96) + 100 \times (57.19 - 46.22) \\
&= 14,369 \\
C &= (2 \times 120 \times 50) + (60 \times 100) + (60 \times 100) + (0.02 \times 440,000) \\
&= 32,800 \\
D &= (0.05 \times 490,000) - (0.02 \times 440,000) \\
&= 15,700
\end{aligned}$$

It can be seen that $B \neq D$ in this case. Then the value of $x = P_{H_0}$ at the point at intersection is calculated.

$$\begin{aligned}
P_{H_0} &= \frac{C - A}{B - D} \\
&= 0.18
\end{aligned}$$

Then we use $P_{H_0} = \frac{C-A}{B-D}$ to find the value of $y = E(\text{total cost})$ at the point of intersection. From (3.10),

$$\begin{aligned} E(\text{total cost}) &= A + B(\frac{C-A}{B-D}); \\ &= \frac{(BA-DA)+(BC-BA)}{B-D}; \\ &= \frac{BC-DA}{B-D} \\ &= 35,571.98. \end{aligned}$$

In this example, $x = \frac{C-A}{B-D}$ lies between 0 and 1, $A \neq C$ and $A > C$. Hence $y_{HNL} > y_{ER}$ for $x < \frac{C-A}{B-D}$ and $y_{HNL} < y_{ER}$ for $x > \frac{C-A}{B-D}$. Figure 3.1 provides an illustration of the expected total costs of the HNL and ER designs. It can be seen that the expected total costs of the two designs have upward trends as P_{H_0} increases. The two designs are equally optimal at $x = \frac{C-A}{B-D} = 0.18$. Initially, for $x < \frac{C-A}{B-D} = 0.18$, the ER design is the optimal design. This is because $A > C$.

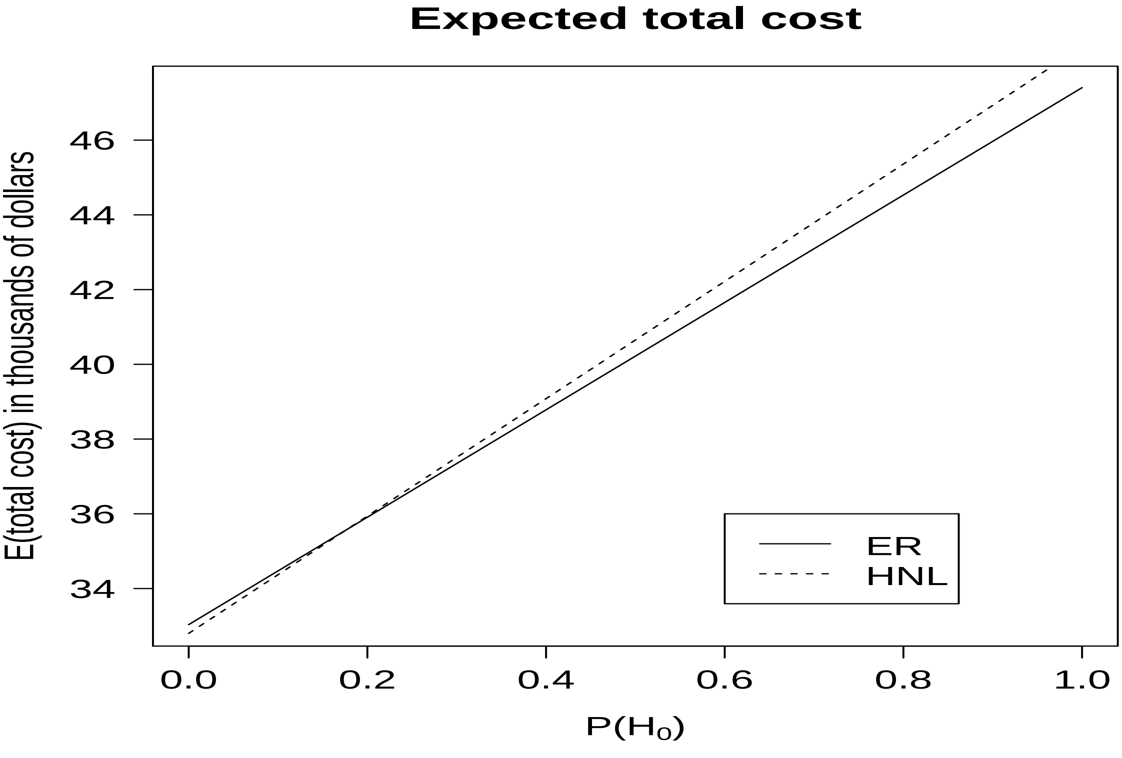


Figure 3.1: E(total cost) of the ER and HNL designs as a function of P_{H_0}.

Let us consider equations (3.10) and (3.12). A and C consist of two main parts: C_{E2} + the cost of all patients (excluding treatment costs) in situation 2 and the cost of lost opportunity when arm B is not selected as a better treatment.

In the first part (C_{E2} + the cost of all patients (excluding treatment costs) in situation 2), this cost for the HNL design is cheaper than that for the ER design. This is because the HNL design requires smaller values of n_2, n_{a2} and n_{b2} than those in the ER design. In contrast, in the second part, the cost for the ER design is cheaper than that for the HNL design because the ER design gives higher power than the HNL design.

In this example, the value of K_1 is considerable higher than those of K_2, K_a and K_b, so A is higher than C.

Now consider B and D in equations (3.10) and (3.12). It can be seen that B and D consist of three parts. The first part is αK_0. Since we control the value of α, α_{ER} and α_{HNL} are identical. Recall that it is supposed that the values of K_0 in the two designs are identical. Hence, the cost of the trial in the first part of B and D are the same. The second part is $(C_{E1} - C_{E2})+$ the difference in cost of all patients (excluding treatment costs) between situations 1 and 2. In this part, the cost of the ER design is zero. So, this cost for the HNL design is higher than that for the ER design. In the ER design, the trial cannot finish early so $n_1 = n_2$, $n_{a1} = n_{a2}$ and $n_{b1} = n_{b2}$. The third part is βK_1. We see that the changes in $E(C_0)$ depend only on the second and third parts. Again as the ER design gives higher power than the HNL design. $\beta_{HNL} K_1$ is higher than $\beta_{ER} K_1$ and the value of K_1 is substantial higher than those of K_2, K_a and K_b, As a result, B is lower than D. If $P_{H_0}(x)$ increases, the influence of B or D on E(total cost) will increase. E(total cost) is based on P_{H_0} multiplied by B or D. This explains why, when $x > \frac{C-A}{B-D} = 0.18$, the HNL design is the optimal design.

In this example, when the values of lost opportunity (such as K_0, K_1) are considerably greater than K_2, K_a and K_b, when P_{H_0} is small, the design with higher power is the optimal design. In contrast, after the point of intersection, the design which uses less resources becomes the optimal design.

3.8.1 Discussion

For the parameters used in this Example, we can see that, from ethical and economic perspectives, the HNL design is the best design, since the resources required by this design are smallest compared to the other two designs. Additionally, in this design, the number of patients allocated to the inferior treatment is small. Although the HNL design gives lower power than the ER design, its power (94.56%) is certainly enough to detect the difference between two arms effectively. In addition, the power obtained from the ER design (97.86%) is very high. It may be desirable to reduce the sample sizes slightly. Conventionally, we use a power of the test just around 90%. The common design is not as effective a design even though it needs smaller resources than the ER design, because it gives significantly lower power (85.74%) than the other two designs.

Chapter 4

Extension of the HNL design

4.1 Introduction

In this chapter, the HNL design will be extended to a design that is applicable to a more realistic situation.

As mentioned in Chapter 1, this thesis focuses on treatment trials. According to Kalish and Begg (1985), the aim of treatment trials is to provide a trial which can compare the efficacy of treatments with precision and validity. In order to achieve this objective, two main factors should be considered: (1) reducing bias and (2) providing an efficient comparison. So far our research has been focusing on the HNL design which only uses response adaptive randomization. This randomization is based only upon the response of the previous patients. The new patient will be favoured to receive the better demonstrating treatment (Biswas and Bhattacharya, 2012). However, it does not recognise the possibility that patients may have some characteristics (or covariates) that might influence the effect of the treatments. Due to this, bias may occur. As discussed above, a good clinical trial should minimize bias and provide an efficient comparison. Hence, covariates should be considered in the randomization procedure.

We thus believed that this is a gap in the HNL design. This is because some prognostic factors might cause the efficient estimation of a treatment effect, but they might also result in the wrong conclusion. For example, a trial detects the difference between two treatments even though there is no difference between these treatments. In fact, the difference is caused by relevant covariates. Thus, this chapter aims to fill a gap in the HNL design by developing an extension of the HNL design. This new design will consider the response of the previous patients and the prognostic factors when allocating a treatment to a new patient. The detail of this design will be described in Section 4.4. In this chapter, firstly, Ning and Huang (2010) will be summarised because we will adopt some methods from this paper to develop the extension of the HNL design. Secondly, A second Ning program (hereafter referred to as "Ning2") will be investigated. The Ning2, which does the calculations for Ning and Huang (2010), was obtained from an email of Huang (one of the authors in NH) by personal communication sent on 16 April 2013. We will also compare the results obtained from the modification of Ning2 with the results shown in Ning and Huang (2010). The detail of the Ning2 and the modified program will be described in Section 4.3. Thirdly, Section 4.4 provides the explanation for why and how the extension of the HNL design is developed. Then the results obtained from the extension of the HNL design will be shown in Section 4.5. Finally, the conclusion will be provided in Section 4.6.

4.2 Ning and Huang (2010)

Ning and Huang (2010) proposed a new design that incorporates the advantages of both the response adaptive (RA) randomization and a covariate-adaptive (CA) randomization. This new design is called a response-adaptive, covariate-adjusted (RACA) randomization design.

The benefit of a RA randomization is that more patients can be allocated to the superior treatment. However, one disadvantage of this randomization is that any differences caused by the covariates are not taken into account. In a clinical trial, the covariates are characteristics of the patients such as gender and age that may affect a response variable. If the trial is conducted without thinking about the covariates, we cannot know whether a detected difference results from the treatments or from the characteristics of the patients.

The aim of covariate-adaptive (CA) randomization is to decide the treatment allocation of a new patient that would achieve the balance of the principal covariates between the treatment groups (Rosenberger and Lachin, 2002).

4.2.1 The RACA design

In this section, we will describe how Ning and Huang (2010) (from now on referred to as NH) constructed the RACA design.

RA randomization

In NH, the RA randomization was carried out by using a Bayesian beta-binomial model for the response. Let Y_{im} be the dichotomous response of the ith patient for treatment m, $m = 0, 1$. Y_{im} can be defined as

$$Y_{im} = \begin{cases} 0 & \text{failure,} \\ 1 & \text{success.} \end{cases}$$

If the covariates are ignored, then the Y_{im} are independently and identically distributed across $i = 1, ..., n_m$. Let s_m denote the probability of success for treatment m, $m = 0, 1$.

Here, it should be noted that NH used p_0 and p_1 to represent the probability of success for treatment 0 (A) and 1 (B) respectively. However, we changed them to s_0 and s_1 to avoid confusion with p_A and p_B. Recall that p_A and p_B were defined as the probability of assigning the current patient to arms A and B respectively.

It was assumed in NH that Y_{im} has a Bernoulli distribution with parameter s_m, where the prior distribution of s_m is a Beta distribution with parameters α_m and β_m.

In this procedure, all covariates were ignored in this assumption. This is because NH did not consider the covariate when making a decision on the superior treatment. Because of this, Bayesian theory can be employed.

Since the Beta distribution is a conjugate prior for s_m in the Bernoulli distribution and the prior distribution of s_m is Beta (α_m, β_m), NH were able to obtain the posterior distribution of s_m. It is Beta $(\alpha_m + n_{m1}, \beta_m + n_{m0})$, where n_{mj} is the number of patients giving response j in treatment arm m; see Mukhopadhyay (2000, p. 482) and Hogg *et al.* (2013, p. 613) for proof. NH assumed that $\alpha_0 = \alpha_1 = \beta_0 = \beta_1 = 1$, expressing reasonably noninformative prior information for the probability of success.

As the trial progressed, the posterior distributions of s_0 and s_1 were continuously updated. Then the probability of allocating the current patient to arm A, $p_A = Pr(s_0 > s_1 \mid data)$, was calculated. This formula was proposed by Thompson (1933). NH used this probability as the criterion for selecting the superior treatment.

For optimal allocation, Rosenberger *et al.* (2001a) and Rosenberger and Hu (2004) argued that the probabilities of assigning the current patient to arm A should be

$$p_{A,RA} = \frac{\sqrt{p_A}}{\sqrt{p_A} + \sqrt{1 - p_A}}. \tag{4.1}$$

It should be noted that the s_m obtained in this procedure was used only for evaluating p_A. It was not used to generate Y_{im}. In each simulation, Y_{im} was generated by using s_{im} obtained from the logistic regression that will be described in the following section.

CA randomization

In this procedure, NH determined how to balance the covariates between two treatments. In NH, three binary covariates were considered: patient age (younger than 60 years, or 60 years and older), cytogenetics (two prognostic categories: favourable or unfavourable), and the number of previous chemotherapy treatments (one, or more than one). This is because NH focused on a cancer trial (e.g. acute myeloid leukemia), and the response of a patient may depend on these covariates.

The impact of the patient on the covariate imbalance was determined when the new patient was recruited. This patient was temporarily assigned to each arm in turn. Then the degrees of covariate imbalance between arms A and B were compared. The details of the method of measuring the degree of covariate imbalance are described in the next section.

For a covariate-adaptive (CA) randomization, the probability of assigning the current patient to treatment A is based on the idea of the biased coin design proposed by Efron (1971). The arm is given a higher probability p_{favour} in the randomization if it minimizes covariate imbalance. This probability is then

$$p_{A,CA} = \begin{cases} p_{favour} & \text{if allocation to A minimizes the imbalance of covariates,} \\ 1 - p_{favour} & \text{if allocation to B minimizes the imbalance of covariates,} \\ 0.5 & \text{if allocation to A or B provides the same imbalance of covariates.} \end{cases} \tag{4.2}$$

p_{favour} can be in the range of more than 0.5 to 1. Pocock (1993) suggested that for a trial that has a sample size less than 100 patients, p_{favour} may be set to 2/3. Otherwise, Pocock (1993) recommended that p_{favour} may be set to 3/4.

However, for this research, the simulations in NH showed that using a p_{favour} of 0.7 or 0.8 gave good performance.

The degree of covariate imbalance

In NH, there were two steps to determine the degree of imbalance of covariate levels. Based on the assumption of equal covariate distributions across treatment arms, the observed numbers of patients in level k of the jth covariate allocated to treatment B should be close to their expected values. For this reason, for each covariate, the metric of the degree of imbalance of a covariate can be defined as

$$D_{jk} = n_{1jk} - (n_{0jk} + n_{1jk})\frac{n_1}{n_0 + n_1}, \tag{4.3}$$

where J is the number of covariates; $j = 1, ..., J$;

k is the level of the relevant covariate; $k = 1, ..., L_j$;

n_m is the number of patients allocated to treatment m;

n_{mjk} is the number of patients belonging to the kth level of the jth covariate in arm m.

Table 4.1: The contingency table for Z_j that shows where the expected value in (4.3) comes from.

k	Treatment		Total
	A	B	
0	n_{0j0}	n_{1j0}	$n_{0j0} + n_{1j0}$
1	n_{0j1}	n_{1j1}	$n_{0j1} + n_{1j1}$
Total	n_0	n_1	$n_0 + n_1$

Table 4.1 illustrates the contingency table for Z_j that shows where the expected value in (4.3) comes from. Under the null hypothesis of no association between the level of the covariate and the treatment arm, the expected frequency in the (2,2) cell is $(n_{0j1} + n_{1j1})n_1/(n_0 + n_1)$.

It should be noted here that when a covariate has only 2 levels ($L_j = 2$), $D_{j0} + D_{j1} = 0$ for each $j = 1, 2, 3$. From (4.3) :

$$D_{j0} = n_{1j0} - (n_{0j0} + n_{1j0})\frac{n_1}{n_0 + n_1},$$

and

$$D_{j1} = n_{1j1} - (n_{0j1} + n_{1j1})\frac{n_1}{n_0 + n_1}.$$

$$D_{j0} + D_{j1} = (n_{1j0} + n_{1j1}) - [(n_{0j0} + n_{0j1}) + (n_{1j0} + n_{1j1})]\frac{n_1}{n_0 + n_1}$$

It can be seen from Table (4.1) that $n_{0j0} + n_{0j1} = n_0$ and $n_{1j0} + n_{1j1} = n_1$. Hence,

$$D_{j0} + D_{j1} = n_1 - (n_0 + n_1)\frac{n_1}{n_0 + n_1} = 0.$$

Consequently, we need to consider only one of D_{j0} or D_{j1} when $L_j = 2$.

At the end of the trial, all metrics are combined across all levels in order to evaluate the level of overall imbalance between the treatment arms. That is,

$$D = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{L_j} |D_{jk}|. \tag{4.4}$$

In NH, there were $J = 3$ covariates. For the CA and RACA designs, after n_0 patients had been enrolled in the trial, each patient was temporarily assigned to each arm in turn as described above. Then D_A and D_B were computed by using formula (4.4) and compared. The current patient would be allocated with a higher probability p_{favour} to the treatment that can minimize covariate imbalance.

However for the ER and RA designs, D was computed only at the end of the trial.

RACA randomization

In order to incorporate the advantages of the CA and RA designs, the assignment of a treatment to a new patient will be determined by both the results of the previous patients and the consideration of the balance of covariates. NH suggested that the probability that a new patient will be assigned to treatment A should be

$$p_{A,RACA} = \frac{p_{A,RA} \cdot p_{A,CA}}{p_{A,RA} \cdot p_{A,CA} + (1 - p_{A,RA})(1 - p_{A,CA})}. \tag{4.5}$$

It should be noted that NH mentioned that $p_{A,RACA}$ was used as the criterion for selecting the superior treatment. In contrast, in the Ning2, Ning used p_A as this criterion. During the modification of the Ning2, we tried in turn to use $p_{A,RACA}$ and p_A as this criterion. Finally, we found that when using p_A as the criterion for selecting the superior treatment, the results obtained from the program were similar to the published results. The details of the Ning2 will be described in Section 4.3.

4.2.2 Simulation

In NH, for a given set of design parameters a total of 5,000 simulations were conducted in order to assess the performance of the RACA design. The properties of this design were compared with those of the ER, RA and CA designs.

Data Generation

NH considered a trial that compared two treatments (A and B) for a dichotomous response (success or failure). In this trial, there was a staggered entry of patients. As mentioned earlier, three binary covariates were considered.

Let Z_1 be the patient's age group. Z_1 can be defined as

$$Z_1 = \begin{cases} 0 & \text{if the patient's age is less than 60 years,} \\ 1 & \text{otherwise.} \end{cases}$$

Let Z_2 be the patient's cytogenetics category. Z_2 can be defined as

$$Z_2 = \begin{cases} 0 & \text{if the prognosis is favorable for the patient,} \\ 1 & \text{if the prognosis is not favorable for the patient.} \end{cases}$$

Let Z_3 be the number of previous chemotherapy treatments. Z_3 can be defined as

$$Z_3 = \begin{cases} 0 & \text{if the patient has been given one chemotherapy treatment,} \\ 1 & \text{more than one.} \end{cases}$$

For each patient, these covariates were independently drawn from Bernoulli distributions with probabilities 0.7, 0.5 and 0.7 respectively.

Model

In NH, due to the binary responses, for each patient, the probability of success was generated from the logistic regression model given by

$$\text{logit}P(Y_i = 1) = \beta_0 + \beta_T T + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3, \tag{4.6}$$

where T is a treatment indicator variable ($T = 0, 1$), Y_i is the outcome of the ith patient, β_0 is the intercept, β_T is the treatment coefficient, β_1 is the coefficient for age, β_2 is the coefficient for cytogenetics, and β_3 is the coefficient for the number of previous chemotherapy treatments.

Thus,

$$P(Y_i = 1) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_T T + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)\}}. \tag{4.7}$$

Although in NH the new design was called a response-adaptive, covariate-adjusted randomization design and used logistic regression to model the probability of success for each patient as in Rosenberger *et al.* (2001b), this new design was different from the covariate-adjusted response-adaptive design mentioned in Rosenberger *et al.* (2001b). According to Rosenberger *et al.* (2001b), the logistic regression model was given by including the treatment-covariate interactions term. In addition, the probability of assigning a new patient to treatment A was based on the estimated covariate odds ratio, defined as the ratio of odds when assigning a new patient to treatment A and when assigning a new patient to treatment B. However NH did not consider the treatment-covariate interactions term in the logistic regression model. This is because considering the interactions term in the regression model is promising only when the sample size is large. Rosenberger *et al.* (2001b) used a sample size of 200 and $n_0 = 85$. On the other hand, NH used sample sizes of 60 and 100 and $n_0 = 10$ and 20 respectively. Unlike Rosenberger *et al.* (2001b), in NH the probability of assigning a new patient to treatment A

was based on both the posterior probability evaluated while the trial progressed, and the degree of covariate imbalance.

Rosenberger and Sverdlov (2008) (hereafter referred to as RS) investigated covariate-adjusted response adaptive (CARA) randomization designs to compare two treatments (A and B) when there is a binary response and covariates. Specifically, they considered a binary and two continuous covariates: gender, age and cholesterol level. The CARA designs were compared to the stratified permuted block design (SPBD), the complete randomization (the ER) designs and the CA designs by considering design characteristics: the probability of Type I error, the statistical power, the total number of treatment failures, the probability of assigning patients to treatment A, the probability of assigning patients to treatment A within the male category of the covariate gender, the Kolmogorov-Smirnov distance between the empirical distributions of covariate age in treatment groups A and B. In stratified permuted block designs (SPBDs), patients are divided into subgroups based on those characteristics that might influence the response. The subgroups are called strata. Randomization is carried out within each stratum by using permuted blocks in order to achieve balance across treatment groups. The aim of a SPBD is to balance on treatment arms by considering the combinations of the covariates.

In RS, the power was defined as the probability of detecting the difference between treatment effects. The seven design characteristics were considered because balance can be measured by the probability of assigning patients to treatment A and the probability of assigning male patients to treatment A. The statistical power can be used to evaluate the efficiency of designs. Moreover, the total number of treatment failures can indicate the ethical property of the design.

It was found that the CARA designs gave similar power to the SPBDs, the CA designs and the ER designs, while CARA designs caused fewer treatment failure than the other designs. Moreover the CARA designs can balance the distributions of the continuous covariates better than the ER designs. Additionally, suppose that arm B is better than arm A. Only the CARA designs had the probability of assigning patients to treatment A less than 0.5. Although, RS considered cholesterol level, they did not give any results for it.

In conclusion, among the SPBDs, the ER designs, the CA designs and the CARA designs, the CARA designs can combine features of a good design covering efficiency and ethics and including balance of both covariates and treatments.

4.3 The Ning2

In this section, we begin by examining the Ning2. As mentioned in Section 4.2.1, we obtained this program from an email of Huang (one of the authors in NH) by personal communication sent on 16 April 2013. We then found that we needed to modify the program. The modification is for two main reasons: (1) some differences between NH and the Ning2; (2) the effectiveness of the Ning2.

4.3.1 Differences

As mentioned above, there are some difference between NH and the Ning2. First of all, in NH, three dichotomous covariates were considered. However, this program was written to consider only one covariate. We therefore modified it to extend this program from considering one covariate to considering three covariates.

Secondly, in the Ning2, the values of β_0, β_1 and β_2 were calculated from the information that is entered into the program. However, in NH, the authors specified these values before carrying out the trial. Hence, these values should be entered directly.

In the modified program, we read these values by specifically asking for the values of β_0, β_1 and β_2.

Finally, as mentioned in Section 4.2.1, the Ning2 used p_A as the criterion for choosing the superior treatment. At the end of the trial, n patients were recruited to the trial. It can be observed from the program that p_A was computed by using the information of only $(n-1)$ patients. On the other hand, NH stated that, at the end of the trial, the comparison for selecting the superior treatment was conducted. We thought that we should use the information from all patients instead of ignoring the information obtained from the last patient. Hence, we evaluated p_A by using the information from all patients.

Note that, in NH, all designs were performed without including an early stop.

4.3.2 Results

We made very minor modifications to the Ning2 program to improve its computational efficiency.

In this section, we will compare the results obtained from the modified program with the published results of NH. NH compared the quality of four designs by comparing five design characteristics: the probability of Type I error, the power of the test, the average number of patients allocated to each arm, the average number of patients who gain a successful outcome from the treatments, and the degree of imbalance of covariate.

As in HNL, in NH, two situations will be considered.

1. In situation 1, there was no prior knowledge of how arms A and B differ.
2. In situation 2, it was supposed that treatment B is superior to treatment A.

In situation 1, the hypotheses were

$H_0 : \mu_a - \mu_b = 0$

$H_1 : \mu_a - \mu_b \neq 0.$

In situation 2, the hypotheses were

$\mathrm{H}_0 : \mu_a \geq \mu_b$;

$\mathrm{H}_1 : \mu_a < \mu_b$.

Hence, in NH, the probability of Type I error was estimated from the proportion of the time that H_0 was rejected in situation 1, whereas the power of the test was estimated from the proportion of the time that H_0 was rejected in situation 2. In other words, the probability of Type I error was estimated from the proportion of the time that arms A or B is selected as a superior treatment in situation 1, whereas the power of the test was estimated from the proportion of the time that B is selected as a superior treatment in situation 2.

We note here that the average number of patients who gain a successful outcome from the treatments (ANPS) is the expected number of patients who get $Y_{im} = 1$.

It should be also noted that ANPS is introduced by us. Actually, NH called this design characteristic the average number of patients who achieved treatment success. In order to make it clear, we changed from the average number of patients who achieved treatment success to the average number of patients who gain a successful outcome from the treatments.

Tables 4.2 and 4.3 compare the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 1, when using $p_U = 0.95$, $N_{max} = 60$, and $n_0 = 10$ and when using $p_U = 0.975$, $N_{max} = 100$, and $n_0 = 20$ respectively. The results displayed in Tables 4.2 and 4.3 illustrate that the probability of Type I error and the average number of patients allocated to each arm obtained from the modified program were similar to those in the published results. However, the ANPS, the standard deviations of the number of patients who gain successful outcome from the treatments (SDNPS) and the degree of covariate imbalance obtained from the modified program were

Table 4.2: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 1 (where the efficacies of arms A and B are equal), when using p_U = 0.95, N_{max} = 60, and n_0 = 10.

Design	Source	Arm	Average (sd)	P(selected)	α	ANPS (sd)	Degree of imbalance
Scenario 1:(β_0 ,β_T, β_1, β_2, β_3) =(0, 0, 1.3, 0.6, 0.4)							
ER	A	A	29.97 (3.88)	0.043	0.092	47.72 (3.17)	0.14
		B	30.03 (3.88)	0.049			
	B	A	29.98 (3.85)	0.049	0.097	34.78 (3.81)	0.29
		B	30.02 (3.85)	0.048			
	C	A	30.06 (3.87)	0.044	0.087	47.73 (3.15)	0.15
		B	29.95 (3.87)	0.043			
CA	A	A	30.05 (5.88)	0.036	0.075	47.70 (3.07)	0.04
		B	29.95 (5.88)	0.039			
	B	A	29.97(4.44)	0.047	0.087	34.77 (3.81)	0.07
		B	30.03 (4.44)	0.040			
	C	A	30.07 (5.76)	0.035	0.070	47.74 (3.11)	0.04
		B	29.93 (5.76)	0.035			
RA	A	A	30.38 (9.64)	0.070	0.126	47.67 (3.17)	0.14
		B	29.62 (9.64)	0.056			
	B	A	29.95 (9.46)	0.065	0.128	34.84 (3.85)	0.28
		B	30.05 (9.46)	0.063			
	C	A	29.88 (9.66)	0.062	0.126	47.70 (3.13)	0.14
		B	30.12 (9.66)	0.064			
RACA	A	A	29.90 (9.55)	0.052	0.111	47.66 (3.13)	0.04
		B	30.10 (9.55)	0.059			
	B	A	30.12 (7.41)	0.045	0.087	34.79 (3.76)	0.08
		B	29.88 (7.41)	0.047			
	C	A	29.85 (9.31)	0.050	0.107	47.66 (3.14)	0.04
		B	30.16 (9.31)	0.057			

Let source A represent the results of the modified program with seed 1234; source B represent the published results of NH; and source C represent the results of the modified program with seed 5678.

Table 4.3: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 1 (where the efficacies of arms A and B are equal), when using $p_U = 0.975$, $N_{max} = 100$, and $n_0 = 20$.

Design	Source	Arm	Average (sd)	P(selected)	α	ANPS (sd)	Degree of imbalance
Scenario 1:(β_0 ,β_T, β_1, β_2, β_3) =(0, 0, 1.3, 0.6, 0.4)							
ER	A	A	50.04 (5.02)	0.025	0.047	79.51 (4.00)	0.11
		B	49.96 (5.02)	0.022			
	B	A	49.98 (5.04)	0.025	0.050	57.88 (4.83)	0.23
		B	50.14 (5.04)	0.025			
	C	A	49.89 (4.95)	0.025	0.050	79.38 (4.05)	0.11
		B	50.11 (4.95)	0.025			
CA	A	A	50.09(7.30)	0.018	0.038	79.50 (4.03)	0.02
		B	49.91 (7.30)	0.020			
	B	A	49.99 (5.80)	0.020	0.041	57.90 (4.87)	0.04
		B	50.01 (5.80)	0.021			
	C	A	49.87 (7.41)	0.017	0.036	79.44 (4.04)	0.02
		B	50.13 (7.41)	0.019			
RA	A	A	49.76 (14.87)	0.033	0.067	79.54 (4.01)	0.11
		B	50.24 (14.87)	0.035			
	B	A	50.19 (14.78)	0.030	0.065	58.00 (4.99)	0.22
		B	49.81 (14.78)	0.035			
	C	A	49.68 (15.14)	0.037	0.074	79.46 (4.09)	0.11
		B	50.33 (15.14)	0.037			
RACA	A	A	50.14 (13.79)	0.026	0.052	79.50 (3.98)	0.02
		B	49.86 (13.79)	0.026			
	B	A	50.07 (11.25)	0.024	0.050	57.87 (4.95)	0.05
		B	49.93 (11.25)	0.026			
	C	A	49.81 (13.99)	0.028	0.057	79.46 (4.02)	0.02
		B	50.19 (13.99)	0.029			

Let source A represent the results of the modified program with seed 1234; source B represent the published results of NH; and source C represent the results of the modified program with seed 5678.

different from those in the published results. For the RACA and CA designs, the standard deviations of the numbers of patients assigned to each arm obtained from the modified program were different from those in the published results. Additionally, the RA design gave the largest Type I error rates. As expected, we obtained the smallest Type I error rates and the lowest degree of imbalance from the CA design. The RA and ER designs gave similar degree of covariate imbalance which is largest. NH worked out the probability of a success for the ER design from equation (4.7) in Section 4.2.2, which is not affected by which allocation method is used. Hence, we can work out the expected probability of success simply by knowing the probability distributions of the three covariates Z_1, Z_2 and Z_3. In NH, these covariates were independently drawn from Bernoulli distributions with probabilities 0.7, 0.5 and 0.7 respectively.

Let $P(Y_i = 1) = g(Z_1, Z_2, Z_3)$.

Then,

$$E(P(Y_i = 1)) = \Sigma_{Z_1}\Sigma_{Z_2}\Sigma_{Z_3} g(Z_1, Z_2, Z_3) P_{Z_1,Z_2,Z_3}(Z_1, Z_2, Z_3)$$

Since Z_1, Z_2 and Z_3 are independent, $P_{Z_1,Z_2,Z_3}(Z_1, Z_2, Z_3) = P(Z_1) \times P(Z_2) \times P(Z_3)$.

Under Scenario 1, (β_0 ,β_T, β_1, β_2, β_3) =(0, 0, 1.3, 0.6, 0.4), the expected probability of a success is 0.7948601. This means that, for 60 patients, we would expect $60 \times 0.7948601 = 47.69$ successes, which agrees very well with our simulated result of 47.72 (for a seed of 1234). Similarly, the SDNPS could be computed from $\sqrt{60 \times 0.7948601 \times (1 - 0.7948601)} = 3.127851$, which corresponds very well to our simulated result of 3.17 (for a seed of 1234). As NH get a quite different result, we suspect that there may be an error in their program or the reported results.

We investigated the effect of altering the values of p_U and thus altering the

Table 4.4: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with seed 1234, in Scenario 1 (where the efficacies of arms A and B are equal), when using different values of p_U, $N_{max} =$ 60, and $n_0 = 10$. The values of p_U have been selected to give a probability of Type I error of approximately 0.10.

Design	Arm	p_U	Average (sd)	P(selected)	α	ANPS (sd)	Degree of imbalance
Scenario 1:(β_0 ,β_T, β_1, β_2, β_3) =(0, 0, 1.3, 0.6, 0.4)							
ER	A	0.945	29.97 (3.88)	0.050	0.107	47.72 (3.17)	0.14
	B		30.03 (3.88)	0.057			
CA	A	0.935	30.05 (5.88)	0.049	0.100	47.70 (3.07)	0.04
	B		29.95 (5.88)	0.051			
RA	A	0.960	30.38 (9.64)	0.059	0.107	47.63 (3.17)	0.14
	B		29.62 (9.64)	0.048			
RACA	A	0.955	29.90 (9.55)	0.047	0.100	47.66 (3.13)	0.04
	B		30.10 (9.55)	0.053			

probability that a treatment is selected as 'better'. Table 4.4 shows the results for the ER, RA, RACA and CA designs obtained from the modified program with seed 1234, in Scenario 1, when using different values of p_U, $N_{max} = 60$, and $n_0 = 10$. The values of p_U have been selected to give a probability of Type I error of approximately 0.10. In addition, Table 4.5 displays the corresponding results for Scenario 1 when using different values of p_U, $N_{max} = 100$, and $n_0 =$ 20. The values of p_U have been selected to give a probability of Type I error of approximately 0.05.

We initially used cut-off values of $p_U = 0.95$ and 0.975, and we obtained Type I error rates as displayed in Tables 4.4 and 4.5. Then various simulations were carried out to find a p_U giving the probability of Type I error around 0.10 and 0.05 respectively. The results shown in Tables 4.4 and 4.5 illustrate that the RA design required the largest values of p_U whereas the CA design needed the smallest values of p_U.

Tables 4.6 and 4.7 compare the results for the ER, RA, RACA and CA designs

Table 4.5: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with seed 1234, in Scenario 1 (where the efficacies of arms A and B are equal), when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05.

Design	Arm	p_U	Average (sd)	P(selected)	α	ANPS (sd)	Degree of imbalance
Scenario 1:(β_0 ,β_T, β_1, β_2, β_3) =(0, 0, 1.3, 0.6, 0.4)							
ER	A B	0.975	50.04 (5.02) 49.96 (5.02)	0.025 0.022	0.047	79.51 (4.00)	0.11
CA	A B	0.966	50.09 (7.30) 49.91 (7.30)	0.025 0.027	0.052	79.50 (4.03)	0.02
RA	A B	0.980	49.76 (14.87) 50.24 (14.87)	0.028 0.028	0.056	79.54 (4.01)	0.11
RACA	A B	0.975	50.14 (13.79) 49.86 (13.79)	0.026 0.026	0.052	79.50 (3.98)	0.02

obtained from the modified program with the published results of NH, in Scenario 2, when using different values of p_U, $N_{max} = 60$ and 100 and $n_0 = 10$ and 20 respectively. In addition, Tables 4.8 and 4.9 show the corresponding results under Scenario 3, when using different values of p_U, $N_{max} = 60$ and 100, and $n_0 = 10$ and 20 respectively.

The results shown in Tables 4.6 - 4.9 illustrate that the average number of patients allocated to each arm obtained from the modified program are similar to those in the published results except for the RA designs shown in Tables 4.7 - 4.9 and except for the RACA design shown in Table 4.9. On the other hand, the power, the ANPS, the SDNPS and the degree of covariate imbalance obtained from the modified program are different from those in the published results. In particular, the degree of covariate imbalance in the published results are consistently twice as great as that obtained from the modified program.

We did not know exactly the values of p_U used by NH. Moreover, the values of p_U strongly influenced the power. This may explain why the power obtained

Table 4.6: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 2 (arm B is better than arm A), when using different values of p_U, $N_{max} = 60$, and $n_0 = 10$.

Design	Source	Arm	p_U	Average (sd)	P(selected)	$1-\beta$	ANPS (sd)	Degree of imbalance
Scenario 2:(β_0 ,β_T, β_1, β_2, β_3) =(0, 1, 1.3, 0.6, 0.4); $\alpha = 0.10$								
ER	A	A	0.945	30.01 (3.85)	0.00	0.34	51.06 (2.78)	0.14
		B		29.99 (3.85)	0.34			
	B	A		29.87 (3.80)	0.00	0.49	40.54 (3.62)	0.29
		B		30.13 (3.80)	0.49			
	C	A	0.945	29.93 (3.87)	0.00	0.33	51.08 (2.73)	0.14
		B		30.06 (3.87)	0.33			
CA	A	A	0.935	30.02 (5.77)	0.00	0.36	51.05 (2.78)	0.04
		B		29.98 (5.77)	0.36			
	B	A		29.83 (4.48)	0.00	0.50	40.52 (3.68)	0.09
		B		30.17 (4.48)	0.50			
	C	A	0.935	30.02(5.76)	0.00	0.36	51.04 (2.81)	0.04
		B		29.98 (5.76)	0.36			
RA	A	A	0.960	21.76 (8.92)	0.00	0.33	52.06 (2.64)	0.13
		B		38.24 (8.92)	0.33			
	B	A		19.62 (8.29)	0.00	0.42	42.56 (3.78)	0.26
		B		40.37 (8.29)	0.42			
	C	A	0.960	21.62(8.76)	0.01	0.32	51.99 (2.69)	0.13
		B		38.38 (8.76)	0.32			
RACA	A	A	0.955	22.73 (9.00)	0.00	0.34	51.91 (2.68)	0.04
		B		37.27 (9.00)	0.34			
	B	A		21.70 (7.20)	0.00	0.50	42.12 (3.66)	0.07
		B		38.29 (7.20)	0.50			
	C	A	0.955	22.87 (8.99)	0.00	0.34	51.84 (2.70)	0.04
		B		37.13 (8.99)	0.34			

Let source A represent the results of the modified program with seed 1234; source B represent the published results of NH; and source C represent the results of the modified program with seed 5678.

Table 4.7: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 2 (arm B is better than arm A), when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$.

Design	Source	Arm	p_U	Average (sd)	P(selected)	$1-\beta$	ANPS (sd)	Degree of imbalance
Scenario 2:(β_0 ,β_T, β_1, β_2, β_3) =(0, 1, 1.3, 0.6, 0.4); $\alpha = 0.05$								
ER	A	A	0.975	50.06 (4.92)	0.00	0.34	85.01 (3.56)	0.11
		B		49.94 (4.92)	0.34			
	B	A		49.90 (5.03)	0.00	0.54	67.48 (4.70)	0.22
		B		50.10 (5.03)	0.54			
	C	A	0.975	49.90 (5.00)	0.00	0.34	85.10 (3.57)	0.11
		B		50.01 (5.00)	0.34			
CA	A	A	0.966	50.00 (7.54)	0.00	0.39	85.17 (3.56)	0.02
		B		50.00 (7.54)	0.39			
	B	A		49.99 (5.80)	0.00	0.58	67.41 (4.68)	0.04
		B		50.01 (5.80)	0.58			
	C	A	0.966	50.10 (7.51)	0.00	0.39	85.15 (3.62)	0.02
		B		49.90 (7.51)	0.39			
RA	A	A	0.980	32.38 (13.00)	0.00	0.35	87.09 (3.41)	0.10
		B		67.62 (13.00)	0.35			
	B	A		28.46 (11.44)	0.00	0.47	71.66 (4.89)	0.20
		B		71.54 (11.44)	0.47			
	C	A	0.980	32.74 (13.13)	0.00	0.34	87.07 (3.46)	0.10
		B		67.26 (13.13)	0.34			
RACA	A	A	0.975	35.16 (12.69)	0.00	0.35	86.74 (3.44)	0.03
		B		64.84 (12.69)	0.35			
	B	A		32.38 (10.41)	0.00	0.53	70.85 (4.78)	0.06
		B		67.62 (10.41)	0.53			
	C	A	0.975	34.59 (12.53)	0.00	0.37	86.96 (3.43)	0.03
		B		65.41 (12.53)	0.37			

Let source A represent the results of the modified program with seed 1234; source B represent the published results of NH; and source C represent the results of the modified program with seed 5678.

Table 4.8: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2), when using different values of p_U, $N_{max} = 60$, and $n_0 = 10$.

Design	Source	Arm	p_U	Average (sd)	P(selected)	$1-\beta$	ANPS (sd)	Degree of imbalance
Scenario 3:(β_0 ,β_T, β_1, β_2, β_3) =(0, 2, 1.3, 0.6, 0.4); α = 0.10								
ER	A	A B	0.945	30.06 (3.90) 29.94 (3.90)	0.00 0.65	0.65	52.79 (2.48)	0.14
	B	A B		29.93 (3.90) 30.07 (3.90)	0.00 0.90	0.90	44.16 (3.50)	0.29
	C	A B	0.945	29.96 (3.93) 30.04 (3.93)	0.00 0.65	0.65	52.73 (2.53)	0.14
CA	A	A B	0.935	30.12 (5.86) 29.89 (5.86)	0.00 0.68	0.68	52.76 (2.60)	0.04
	B	A B		29.97 (4.44) 30.03 (4.44)	0.00 0.92	0.92	44.26 (3.43)	0.08
	C	A B	0.935	30.12 (5.82) 29.89 (5.82)	0.00 0.68	0.68	52.69 (2.62)	0.04
RA	A	A B	0.960	17.09 (7.50) 42.92 (7.50)	0.00 0.64	0.64	54.90 (2.05)	0.13
	B	A B		13.22 (6.00) 46.77 (6.00)	0.00 0.83	0.83	46.77 (3.07)	0.08
	C	A B	0.960	17.13 (7.32) 42.87 (7.32)	0.00 0.63	0.63	54.94 (2.05)	0.13
RACA	A	A B	0.955	18.62 (8.02) 41.38 (8.02)	0.00 0.66	0.66	54.62 (2.18)	0.05
	B	A B		15.56 (6.25) 44.44 (6.25)	0.00 0.89	0.89	48.80 (3.10)	0.13
	C	A B	0.955	18.89 (7.95) 41.11 (7.95)	0.00 0.65	0.65	54.64 (2.17)	0.05

Let source A represent the results of the modified program with seed 1234; source B represent the published results of NH; and source C represent the results of the modified program with seed 5678.

Table 4.9: Comparison of the results for the ER, RA, RACA and CA designs obtained from the modified program with the published results of NH, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2), when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$.

Design	Source	Arm	p_U	Average (sd)	P(selected)	$1-\beta$	ANPS (sd)	Degree of imbalance
Scenario 3:(β_0 ,β_T, β_1, β_2, β_3) =(0, 2, 1.3, 0.6, 0.4); $\alpha = 0.05$								
ER	A	A	0.975	49.94 (4.99)	0.00	0.76	87.90 (3.28)	0.11
		B		50.06 (4.99)	0.76			
	B	A		49.93 (4.96)	0.00	0.97	73.78 (4.30)	0.22
		B		50.07 (4.96)	0.97			
	C	A	0.975	49.95 (4.95)	0.00	0.76	87.90 (3.27)	0.11
		B		50.05 (4.95)	0.76			
CA	A	A	0.966	49.94 (7.42)	0.00	0.80	87.88 (3.35)	0.02
		B		50.06 (7.42)	0.80			
	B	A		49.94 (5.84)	0.00	0.97	72.75 (4.46)	0.04
		B		50.06 (5.84)	0.97			
	C	A	0.966	49.95(7.32)	0.00	0.80	87.91 (3.40)	0.02
		B		50.05 (7.32)	0.80			
RA	A	A	0.980	23.69 (9.62)	0.00	0.73	92.32 (3.77)	0.10
		B		76.30 (9.62)	0.73			
	B	A		18.06 (6.87)	0.00	0.92	83.79 (3.77)	0.17
		B		81.94 (6.87)	0.92			
	C	A	0.980	23.87 (9.76)	0.00	0.72	92.28 (2.55)	0.09
		B		76.13 (9.76)	0.72			
RACA	A	A	0.975	26.23 (10.33)	0.00	0.77	91.87 (2.65)	0.03
		B		73.77 (10.33)	0.77			
	B	A		20.85 (7.67)	0.00	0.96	82.91 (3.92)	0.08
		B		79.15 (7.67)	0.96			
	C	A	0.975	26.52 (10.58)	0.00	0.77	91.88 (2.70)	0.04
		B		73.48 (10.58)	0.77			

Let source A represent the results of the modified program with seed 1234; source B represent the published results of NH; and source C represent the results of the modified program with seed 5678.

from the modified program was different from that in the published results.

As mentioned in Section 4.3.1, there are some difference between NH and the Ning2. Furthermore, this Ning2 was written to consider only one covariate whereas in NH, three covariates were considered. Hence, we cannot ensure that the method for computing the degree of covariate imbalance in the modified program based on NH is the same as the method used in the Ning2. This may explain why the degree of covariate imbalance obtained from the modified program was different from that in the published results.

Finally, we are not convinced that the errors are caused by us because there are other errors in the Ning2.

4.4 Extension of the HNL design

In this section, we will explain why and how we develop an extension of the HNL design.

NH considered the RACA design for comparing two treatments (A and B) with a binary response. We seek to extend this work by posing the question of how to develop this design for a continuous response.

Moreover, NH's new design was conducted without determining whether to stop the trial early. The benefits of early termination of the trial are for ethical and economic reasons. Hence, we thought that the trial should be carried out by including the possibity of an early stop.

In addition, we found a gap in the HNL design, namely that this design was constructed without considering any prognostic factors. During the planning of a clinical trial, one factor that we need to think about is the validity of the design. Chang (2008, p. 330) stated that, for internal validity, confounding variables should be eliminated. HNL did not consider the characteristics of the patients such as age and the number of previous chemotherapy treatments in the study.

However, it is evident that these two characteristics are possible confounding variables for the response of the patient (NH). Hence the HNL study is limited. We thought that we cannot eliminate confounding variables in our study but we should try to balance their affect across treatments. In order to cope with this problem, we consider the confounding variables as covariates.

For the three reasons mentioned above, we develop an extension of the HNL design for comparing two treatments with a continuous response. In this design, the probability of assigning a current patient is based not only on the response from the previous patients but on the degree of covariate imbalance as well. We also include an interim analysis to determine whether to stop the trial early for the efficacy. The detail of the extension of the HNL design will be described in the following section.

4.4.1 RA randomization

Recall that in Section 1.6.1, $T_{x,i}$ was defined as the progression-free survival time of participant i in treatment x. If we ignore the covariates, then the $T_{x,i}$ are independently and identically distributed across $i = 1, ..., n_x$. In HNL, conditional upon belonging to the kth category of a short-term response, $T_{x,i}$ is assumed to have an exponential distribution with rate $\lambda_{x,k}$.

Following NH, in this section, all covariates are ignored in this assumption. This is because they are not considered when deciding on the superior treatment and stopping trial early. Due to this, we can employ Bayesian theory as in HNL.

In this section, we follow HNL; see the details in Section 1.6.1. We also adopt this procedure from the estimation procedure of HNL; see the details in Section 2.1.1. However, there is a difference in the probability of assigning a patient to arm A. In HNL, a subsequent patient was allocated to treatment A with probability p_A. On the other hand, in this subsection, following Rosenberger *et al.* (2001a) and Rosenberger and Hu (2004), the probability of assigning the current patient to arm A, $p_{A,RA}$, is given by (4.1).

Following HNL, for the RA and RACA designs, if $p_A > p_U$ (or $p_B < p_L$), treatment A (or B) is chosen as the better treatment and the trial will be terminated. These comparisons will be considered every week after the initial patients are enrolled. Then we will conduct the comparisions once more at the end of the trial. The study cannot be terminated early once the recruitment period has finished. The principal aim of the evaluation of p_A is to assign with higher probability the current patient to the treatment showing more efficacy from the accumulated information of the previous patients. In the follow up period, all patients have been entered into the trial. Hence, it is not necessary to evaluate p_A. It would also be possible to update and evaluate p_A after the last patient is enrolled.

For the ER and CA designs, we use the same criteria as for the RA and RACA designs. However, the determination will be performed only at the end of the trials. Consequently, the ER and CA designs cannot stop early.

4.4.2 CA randomization

In a clinical trial, one aim of a good design is to reduce variability. In particular, biases occur from the imbalance of important prognostic factors. In order to overcome these biases, stratified randomization and the minimization method have been addressed.

Many papers such as Pocock and Simon (1975) and Hagino *et al.* (2004) argued that one drawback of stratified randomization is that it is not an appropriate approach for the case that has many prognostic factors. In particular, the sample is usually too small .

In contrast, the minimization approach is more flexible than the stratified randomization in the situation mentioned above. The minimization method was introduced by Taves (1974). In this method, for each treatment the number of previous patients belonging in the same levels of the covariates as a new patient will be determined separately. Next these numbers are combined over all factors. The new patient will be assigned to a treatment giving a lower degree of covariate imbalance. Taves (1974) only determined the number of previous patients belonging in each level of the covariates. However, he did not take into account the new patient.

Barbachano *et al.* (2008) argued that one disadvantages of Taves' method is that we may encounter a problem of predictability. This problem can be defined as being able to predict a treatment which will be received by a subsequent patient by using the information of the previous patients when knowing his/her characteristics. This problem may occur because, in Taves' method, $p_{favour} = 1$. Hence, the idea of the biased coin design proposed by Efron (1971) was used by Pocock and Simon (1975). As mentioned in Section 4.2.1, using the idea of the biased coin design, p_{favour} may range from more than 0.5 to 1. Therefore, it may reduce the problem of predictability.

Hu *et al.* (2014) very recently reviewed techniques for controlling covariates in clinical trial designs such as stratified and covariate-adaptive randomizations. The benefits and disadvantages of these two randomizations were also compared on page 112. A minimization approach was described and discussed. We have used the minimization approach in Section 4.4.2.

Hu *et al.* (2014) recommended the use of the CA designs if balance on covariates is of concern.

Frane (1998) determined the degree of imbalance of covariate levels by comparing the values of the chi-squared goodness-of-fit test statistics for each of the categorical covariates.

For example, suppose that there are two covariates in the trial: smoking and gender. We take into consideration the smoking habits and genders of the people already in the trial. Let us consider the case of a male smoker who is assigned to arm A. The chi-squared test statistics for both covariates will have been determined. For arm A, suppose that the chi-squared values for smoking and gender are 3 and 2 respectively. The covariate with the higher chi-squared statistic, 3, will be chosen.

Table 4.10 shows an example of how to obtain this statistics in the case of a male smoker who is assigned to arm A. By using the chi-squared goodness-of-fit test, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ where the expected frequency (E_i) for smoking is $0.5 \times 12 = 6$. Hence, the chi-squared statistics shown in Table 4.10 were obtained.

Table 4.10: An example of how to obtain this statistic in the case of a male smoker who is assigned to arm A

Covariate		Number of patients		Chi-squared statistics
		A	B	
Smoking	yes	3	9	3
Gender	male	6	2	2

Now let the same process be carried out for arm B. Suppose that if this patient is assigned to arm B, the higher chi-squared statistic is 4. Because the patient should be allocated to the treatment that minimizes the values of the chi-squared test statistics, he should be assigned to arm A. Table 4.11 illustrates an example of assigning a patient by using the Frane method when there are two covariates.

Table 4.11: An example of assigning a patient by using the Frane method when there are two covariates

Arm	Chi-squared statistics		
	Smoking	Gender	higher
A	3	2	3
B	1	4	4

According to Frane (1998) and NH, the method used by them to determine the degree of imbalance of covariate levels for the categorical covariate is based on the difference between the observed and expected numbers of patients. In this research, we consider the three covariates as in NH. Since these covariates are categorical, the degree of imbalance of covariate levels will be determined by using the difference between the observed and expected numbers of patients as well.

However, in both papers, the determination of imbalance of each covariate was performed *separately* for each covariate. Obviously, in real life, the combination of patients' characteristics affects a response variable simultaneously. Consequently, the covariates should be considered in conjunction with one another. If we have two covariates, suppose that 20 patients receive treatment A, and 10 have $(Z_1 = 0, Z_2 = 1)$ and 10 have $(Z_1 = 1, Z_2 = 0)$. Also suppose that 20 patients receive treatment B and 10 have $(Z_1 = 0, Z_2 = 0)$ and 10 have $(Z_1 = 1, Z_2 = 1)$. If we measure the covariate imbalance by looking at each covariate separately such as in NH, it can be concluded that there is no covariate imbalance. This is because, in both treatments, the numbers of patients having $Z_1 = 0, Z_1 = 1, Z_2 = 0$ and $Z_2 = 1$ are identical. In contrast, by looking at these covariates in pairs, in the situation mentioned above, the covariate levels in the two treatments are not balanced. If the sample of 20 patients in a particular arm is balanced, we should have five each of $(Z_1 = 0, Z_2 = 0)$, $(Z_1 = 0, Z_2 = 1)$, $(Z_1 = 1, Z_2 = 0)$ and $(Z_1 = 1, Z_2 = 1)$.

If there are J covariates, there are $\binom{J}{2}$ pairs of covariates: (1,2),..., $(J-1, J)$, so we would have a sum $\sum_{j_1=1}^{J-1} \sum_{j_2=j_1+1}^{J} |D_{j_1 j_2}|$ over pairs of covariates. The number of combinations of levels for covariates j_1 and j_2 is $L_{j_1} \times L_{j_2}$. The quantity $|D_{j_1 j_2}|$ measures the degree of imbalance of covariates taking all levels of any covariates into consideration. This will be defined later in this section.

Following NH, we have three covariates (Z_1, Z_2, Z_3). Then when looking at them in pairs, we have $\binom{3}{2} = 3$ pairs, that is (Z_1, Z_2), (Z_2, Z_3) and (Z_1, Z_3). In addition, in each covariate, there are two levels. Thus for each pair, we have $2 \times 2 = 4$ levels. We create three (4×2) contingency tables. An example of the contingency table for (Z_1, Z_2) is in Table 4.12.

Table 4.12: The contingency table for (Z_1, Z_2)

$(Z_1 = i, Z_2 = m)$	Treatment		Total
	A	B	
$Z_1 = 0, Z_2 = 0$			n_{11}
$Z_1 = 0, Z_2 = 1$			n_{12}
$Z_1 = 1, Z_2 = 0$			n_{13}
$Z_1 = 1, Z_2 = 1$			n_{14}
Total	n_A	n_B	$n_A + n_B$

The remaining tables for the other covariate pairs follow similarly.

It is possible to look at the covariates in triples. We suppose that there are 40 patients allocated to treatment A and 40 patients allocated to treatment B. It is supposed also that the numbers of patients in each combination of levels of covariates in the two arms are as in Table 4.13.

Table 4.13: An example of the numbers of patients in each combination of levels of covariates in the two arms when there are triple covariates

Treatment		$Z_2 = 0$		$Z_2 = 1$	
		$Z_3 = 0$	$Z_3 = 1$	$Z_3 = 0$	$Z_3 = 1$
A	$Z_1 = 0$	10	0	0	10
	$Z_1 = 1$	0	10	10	0
B	$Z_1 = 0$	0	10	10	0
	$Z_1 = 1$	10	0	0	10

If the imbalance of covariates is determined by looking at each covariate separately or at pairs of covariates, we would conclude that there is no covariate imbalance. For example, using the method of measuring covariate imbalance proposed by NH, it can be seen that the numbers of patients having $Z_1 = 0, Z_1 = 1, Z_2 = 0,$

$Z_2 = 1$, $Z_3 = 0$ and $Z_3 = 1$ in the two arms are equal. In addition, when looking at covariates in pairs, we have ten patients in each combination of levels of covariate pairs for the sample of 40 patients in one arm. However, by considering the covariates in triples, the combinations of levels of covariates triples are imbalanced. If the sample of 40 patients for one arm is balanced, the numbers of patients having each of the combinations of levels of covariates triples for that arm should be five.

So clearly the best methods for examining the imbalance of covariates will be to look at the J-way table of covariates. However, for conciseness there will only be a consideration of covariates in pairs.

Note that although Frane (1998) used the chi-squared goodness-of-fit test to measure the degree of covariate imbalance, in this research we consider the method of the chi-squared test for independence. We have two categorical variables; that is, treatment and pairs of level of covariates. Our aim is to investigate whether the former variable is independent of the latter variable. In addition, we see that the expected numbers in (4.3) are based on the formula in the chi-squared test for independence. This may explain why the method of the chi-squared test for independence is used to determine the degree of imbalance of covariate levels.

Let p denote the number of the covariate pair under consideration: $p = 1, ..., \binom{J}{2}$;
and let c denote the particular combination of levels for covariates j_1 and j_2: $c = 1, ..., L_{j_1} \times L_{j_2}$.

As in NH, in order to balance the effect of the covariate in the two treatments, we consider the degree of covariate imbalance by using the assumption of equal covariate distributions across treatment arms. Thus, we compute the differences between the observed numbers of patients in the various levels of the covariate pairs allocated to treatment B and their expected numbers if balance exists.

The metric of the degree of imbalance of covariate is given by

$$D_{pc} = n_{Bpc} - E_{Bpc}, \tag{4.8}$$

where n_{Bpc} is the observed numbers of patients in the cth combination of the pth covariate pair allocated to treatment B; E_{Bpc} is the expected number of patients in the cth combination of the pth covariate pair allocated to treatment B; $E_{Bpc} = (n_{pc} \times n_B)/(n_A + n_B)$; n_{pc} is the number of patients in level c of the pth covariate pair; n_A is the number of patients assigned to treatment A; n_B is the number of patients assigned to treatment B. The level of overall imbalance between the treatment arms is then

$$D = \frac{1}{n} \sum_{p=1}^{\binom{J}{2}} \sum_{c=1}^{L_{j_1} \times L_{j_2}} |D_{pc}|, \tag{4.9}$$

where n is the numbers of patients used in the trial.

For the CA and RACA designs, after the initial patients (patients who are assigned to the treatments by using equal randomization) are enrolled, the CA randomization is commenced. When each new patient is recruited, he/she will be tentatively assigned to both arms in turn to compare the degree of covariate imbalance. Then we give a higher probability $p_{favour} = 0.8$ to the arm that can minimize covariate imbalance. As in NH, by using probability $p_{favour} = 0.8$ or 0.7 in the simulation, the results obtained are satisfactory. Thus, following NH the probability of covariate-adaptive (CA) randomization to treatment A is

$$p_{A,CA} = \begin{cases} p_{favour} & \text{if } D_A > D_B, \\ 1 - p_{favour} & \text{if } D_A < D_B, \\ 0.5 & \text{if } D_A = D_B. \end{cases} \tag{4.10}$$

For the ER and RA designs, (4.9) will be calculated only at the end of the trials.

In HNL, $N_{max} = 120$, $n_0 = 1$ and 30. Similarly, in this chapter, we use $N_{max} = 120$. However, we only use $n_0 = 30$. In HNL, they only focused on the RA design. In contrast, this chapter considers not only the RA design but the ER, CA, and RACA designs as well. One aim of the CA and RACA designs is to balance principal covariates across all treatments. In order to determine the degree of covariate imbalance when assigning the current patient to arms A or B, we need to recruit a group of initial patients. It can be seen that having only one initial patient is not enough to consider the degree of covariate imbalance.

4.4.3 RACA randomization

We adopt (4.5) of this thesis from NH; see the details in Section 4.2.1.

4.4.4 Simulation

The evaluation of the performance of the RACA design was conducted using 5,000 simulations for each set of design parameters. We also compared the quality of this design with that of the ER, RA and CA designs. Again the design characteristics mentioned in Section 3.3 were employed as criteria to assess the four designs. Additionally, in the present analysis, the degree of covariate imbalance was used as a criterion to compare the designs.

In this research, two situations are considered.

- In situation 1, there is no prior knowledge of how arms A and B differ.
- In situation 2, treatment B is supposed to be superior to treatment A.

Again, in these situations, the hypotheses will be

$\mathrm{H}_0 : \mu_a \geq \mu_b$;

$\mathrm{H}_1 : \mu_a < \mu_b$.

It should be noted here that from now on in this chapter we focus only upon this one-sided alternative hypothesis. This is different from NH. In NH, for the first situation, a two-sided hypothesis was considered. On the other hand, for the second situation, NH considered this one-sided hypothesis. Unlike NH, in this research, the Type I error will be estimated from the proportion of the time that arm B is selected as a superior treatment in situation 1.

Note that although RS considered the same two situations as in this research, they did not use the probabilities $p_{x,k}$. Hence, they were considering a simpler situation.

Data Generation

In the data generation, we will simulate the category variable and the progression-free survival time of each patient.

Firstly, in the data generation, for each patient, the category variable is generated. As described in Section 1.6.1, in this research, when patient i in arm x belongs in the kth category of a short-term response, this is represented by $S_{x,k,i} = 1$ and $S_{x,j,i} = 0$ for $1 \leq j \leq 4, j \neq k$. The vectors $(S_{x,1,i}, ..., S_{x,4,i})$ are assumed to be i.i.d. across $i = 1, ..., n_x$ and to have a multinomial $(1, p_{x,1}, ..., p_{x,4})$ distribution.

In order to avoid confusion with the $p_{x,k}$ for estimation procedure, for $p_{x,k}$ used to generate, we will denote by $\pi_{x,k}$ the value that was given for this parameter for each scenario (e.g. under Scenario 1, $\pi_{x,1} = 0.2, \pi_{x,2} = 0.4, \pi_{x,3} = 0.1, \pi_{x,4} = 0.3$). Thus $(S_{x,1,i}, S_{x,2,i}, S_{x,3,i}, S_{x,4,i})$ for all patients was drawn from a multinomial $(1, \pi_{x,1}, \pi_{x,2}, \pi_{x,3}, \pi_{x,4})$ distribution.

Recall that HNL defined $T_{x,i}$ as the progression-free survival time of participant i in arm x. Conditional on occupying the kth category of a short-term response, $T_{x,i}$ is assumed to have an exponential distribution with rate $\lambda_{x,k}$. Therefore, $T_{x,i} \sim \sum_{k=1}^{4} p_{x,k}\text{Exp}(\lambda_{x,k})$. After we calculate $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$ from (4.12), $T_{x,i}$ is simulated from an exponential $\lambda_{x,k}$ distribution.

It should be noted that in a given set of simulations, $\pi_{x,k}$ is kept constant.

Since in this data generation procedure, $p_{x,k} = \pi_{x,k}$, $p_{x,k}$ also is kept constant. Hence, $(S_{x,1,i}, ..., S_{x,4,i})$ is i.i.d. across $i = 1, ..., n_x$. In contrast, the $p_{x,k}$ in the RA randomization are drawn from Dir $(\gamma_{x,k} + n_{x,k})$.

It should be noted also that in each simulation, $T_{x,i}$ is generated by using $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$ obtained from a Generalised Linear Model that will be described in the next section. However, the $\mu_{x,k}$ in the RA randomization is drawn from IG $(\alpha_{x,k} + \sum_{i=1}^{n_{x,k}} \delta_{x,i}^{(k)}, \beta_{x,k} + \sum_{i=1}^{n_{x,k}} t_{x,i}^{(k)})$.

Model

In this model, the three binary covariates of NH will be considered. It is assumed that $Z_1 \sim$ Bernoulli(0.7) as in NH. Unlike NH, we assume that $Z_2 \sim$ Bernoulli and $Z_3 \sim$ Bernoulli with *conditional* probabilities given in Tables 4.14 and 4.15 respectively. The distributions of Z_2 and Z_3 are conditional on the value of Z_1. In reality, it is rare to find independent covariates. Some covariates may depend upon another. For example, if the patient's age is more than 60 years, he/she may have a higher probability of having been given more than one chemotherapy treatment than only one treatment.

Table 4.14: The conditional probabilities for Z_2

j	i	$P(Z_2 = i \mid Z_1 = j)$
0	0	0.60
	1	0.40
1	0	0.35
	1	0.65

$T_{x,i}^{(k)}$ can be defined as the progression-free survival time of patient i if she/he is in category k and arm x. Since $T_{x,i}^{(k)} \sim \text{Exp}(\lambda_{x,k})$ and $\text{E}(T_{x,i}^{(k)}) = \frac{1}{\lambda_{x,k}}$, we can model $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$ by using a Generalised Linear Model. The link function used is the log link. Although the canonical link function for the exponential distribution is the reciprocal function, Myers *et al.* (2010, p. 215) argued that

Table 4.15: The conditional probabilities for Z_3

j	i	$P(Z_3 = i \mid Z_1 = j)$
0	0	0.5
	1	0.5
1	0	0.4
	1	0.6

using the reciprocal link may result in negative values of the response. In the exponential distribution model, the response values are nonnegative. In order to avoid this problem, in this research the log link was chosen as the link function. Let $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$ denote the mean progression-free survival time of the kth category in arm x. This can be based on the following model:

$$\log[\mu_{x,k}] = -\log[\lambda_{x,k}] = \beta_{0k} + \beta_{Tk}T + \beta_{1k}Z_1 + \beta_{2k}Z_2 + \beta_{3k}Z_3 \tag{4.11}$$

where T is a binary treatment variable. We defined T to be an indicator variable

$$T = \begin{cases} 0 & \text{if patient is allocated to treatment A,} \\ 1 & \text{if patient is allocated to treatment B.} \end{cases}$$

Then,

$$\mu_{x,k} = \frac{1}{\lambda_{x,k}} = \exp(\beta_{0k} + \beta_{Tk}T + \beta_{1k}Z_1 + \beta_{2k}Z_2 + \beta_{3k}Z_3). \tag{4.12}$$

The quantity k is a category variable: $k = 1, 2, 3, 4$, β_{0k} is the intercept of the kth category, β_{Tk} is the treatment coefficient of the kth category, β_{1k} the coefficient for age of the kth category, β_{2k} the coefficient for cytogenetics of the kth category, and β_{3k} the coefficient for the number of previous chemotherapy treatments of the kth category.

In reality, either arm may be superior. However, for simplicity in the simulations, arm B is assumed to be the superior treatment with longer mean progression-free survival time if arms A and B are not identical. In Scenarios

1 and 2, there is no treatment effect on the mean progression-free survival time of the kth category in arm x, so $\beta_{Tk} = 0$.

Note that, in Scenario 2, although the mean progression-free survival times of the kth category are identical in both arms, arm B is the superior treatment. This is because, in this scenario, arm B is assumed to have higher partial PR and CR probabilities (i.e. higher probabilities for categories 3 and 4).

Table 4.16: The values of β_{0k}, β_{Tk}, β_{1k}, β_{2k} and β_{3k}

Scenario	k	β_{Tk}	β_{0k}	β_{1k}	β_{2k}	β_{3k}
1	1	0	1.65	-0.20	-0.17	-0.20
	2	0	3.80	-0.19	-0.23	-0.20
	3	0	4.60	-0.15	-0.20	-0.10
	4	0	4.90	-0.15	-0.10	-0.10
2	1	0	1.65	-0.20	-0.17	-0.20
	2	0	3.80	-0.19	-0.23	-0.20
	3	0	4.60	-0.15	-0.20	-0.10
	4	0	4.90	-0.15	-0.10	-0.10
3	1	0.50	1.65	-0.20	-0.17	-0.20
	2	0.35	3.80	-0.19	-0.23	-0.20
	3	0.38	4.60	-0.15	-0.20	-0.10
	4	0.38	4.90	-0.15	-0.10	-0.10

Table 4.16 shows the values of β_{0k}, β_{Tk}, β_{1k}, β_{2k} and β_{3k}. Additionally, Table 4.17 shows the values of $\mu_{x,k}$ in Scenarios 1 and 2 for all possible covariate combinations when k = 1, 2, 3, 4 when the parameters in Table 4.16 are used. We are trying to generate a variety of situations for possible values of the covariates. We are also attempting to generate situations where the mean of the survival time is similar to the values used by HNL; as a result, we have chosen the values shown in Tables 4.16 and 4.17.

Tables 4.18 - 4.20 show the values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when k = 1, 2, 3, 4, in Scenarios 1 - 3 respectively when the parameters in Table 4.16 are used.

Table 4.17: The values of $\mu_{x,k}$ in Scenarios 1 and 2 for all possible covariate combinations when k = 1, 2, 3, 4 when the parameters in Table 4.16 are used.

k	Z_1	Z_2	Z_3	$\log(\mu_{x,k})$	$\mu_{x,k}$	HNL
1	0	0	0	1.65	5.21	4
	1	1	1	1.08	2.94	
	1	0	0	1.45	4.26	
	0	1	0	1.48	4.39	
	0	0	1	1.45	4.26	
	1	1	0	1.28	3.60	
	0	1	1	1.28	3.60	
	1	0	1	1.25	3.49	
2	0	0	0	3.80	44.70	30
	1	1	1	3.18	24.05	
	1	0	0	3.61	36.97	
	0	1	0	3.57	35.52	
	0	0	1	3.60	36.60	
	1	1	0	3.38	29.37	
	0	1	1	3.37	29.08	
	1	0	1	3.41	30.27	
3	0	0	0	4.60	99.48	75
	1	1	1	4.15	63.43	
	1	0	0	4.45	85.63	
	0	1	0	4.40	81.45	
	0	0	1	4.50	90.02	
	1	1	0	4.25	70.11	
	0	1	1	4.3	73.70	
	1	0	1	4.35	77.48	
4	0	0	0	4.90	134.29	110
	1	1	1	4.55	94.63	
	1	0	0	4.75	115.58	
	0	1	0	4.80	121.51	
	0	0	1	4.80	121.51	
	1	1	0	4.65	104.58	
	0	1	1	4.70	109.95	
	1	0	1	4.65	104.58	

Table 4.18: The values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when $k = 1, 2, 3, 4$, in Scenario 1 when the parameters in Table 4.16 are used.

Arm	p_1	p_2	p_3	p_4	Z_1	Z_2	Z_3	μ_1	μ_2	μ_3	μ_4
A	0.2	0.4	0.1	0.3	0	0	0	5.21	44.70	99.48	134.29
					1	1	1	2.94	24.05	63.43	94.63
					1	0	0	4.26	36.97	85.63	115.58
					0	1	0	4.39	35.52	81.45	121.51
					0	0	1	4.26	36.60	90.02	121.51
					1	1	0	3.60	29.37	70.11	104.58
					0	1	1	3.60	29.08	73.70	109.95
					1	0	1	3.49	30.27	77.48	104.58
B	0.2	0.4	0.1	0.3	0	0	0	5.21	44.70	99.48	134.29
					1	1	1	2.94	24.05	63.43	94.63
					1	0	0	4.26	36.97	85.63	115.58
					0	1	0	4.39	35.52	81.45	121.51
					0	0	1	4.26	36.60	90.02	121.51
					1	1	0	3.60	29.37	70.11	104.58
					0	1	1	3.60	29.08	73.70	109.95
					1	0	1	3.49	30.27	77.48	104.58

Table 4.19: The values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when $k = 1, 2, 3, 4$, in Scenario 2 when the parameters in Table 4.16 are used.

Arm	p_1	p_2	p_3	p_4	Z_1	Z_2	Z_3	μ_1	μ_2	μ_3	μ_4
A	0.2	0.4	0.1	0.3	0	0	0	5.21	44.70	99.48	134.29
					1	1	1	2.94	24.05	63.43	94.63
					1	0	0	4.26	36.97	85.63	115.58
					0	1	0	4.39	35.52	81.45	121.51
					0	0	1	4.26	36.60	90.02	121.51
					1	1	0	3.60	29.37	70.11	104.58
					0	1	1	3.60	29.08	73.70	109.95
					1	0	1	3.49	30.27	77.48	104.58
B	0.1	0.1	0.2	0.6	0	0	0	5.21	44.70	99.48	134.29
					1	1	1	2.94	24.05	63.43	94.63
					1	0	0	4.26	36.97	85.63	115.58
					0	1	0	4.39	35.52	81.45	121.51
					0	0	1	4.26	36.60	90.02	121.51
					1	1	0	3.60	29.37	70.11	104.58
					0	1	1	3.60	29.08	73.70	109.95
					1	0	1	3.49	30.27	77.48	104.58

4.5 Results

It should be noted that here we use the same abbreviations as in Section 3.3. In each simulation, data were generated by using a given set of design parameters shown in Tables 4.16 - 4.20.

Table 4.21 shows the results for the ER, RA, RACA and CA designs, in Scenario 1, when using $p_U = 0.95$, $N_{max} = 100$, and $n_0 = 20$. In addition, Tables 4.22 and 4.23 show the corresponding results, in Scenario 1, when using $p_U = 0.95$, $n_0 = 30$, and $N_{max} = 120$ and 130 respectively. The results displayed in Tables 4.21- 4.23 illustrate that, when using $p_U = 0.95$, the RA design gave the largest degree of covariate imbalance. As expected, we obtained the lowest degree of imbalance from the CA design. The ANP increased as N_{max} became larger. On the other hand, the degree of covariate imbalance decreased as N_{max} increased. The reasons will be provided on page 133.

Table 4.24 shows the results for the ER, RA, RACA and CA designs in Scenario 1 when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. Tables 4.25 and 4.26 show the corresponding results, also in Scenario 1, when using different values of p_U, $n_0 = 30$, and $N_{max} = 120$ and 130 respectively.

Initially, we used a cut-off value of $p_U = 0.95$, and we obtained Type I error rates as shown in Tables 4.21-4.23. Then we performed simulations to find a p_U that gives a the probability of Type I error around 0.05. The results displayed in Tables 4.24-4.26 show that the RACA and RA designs required larger values of p_U than the ER and CA designs. By using a cut-off value of $p_U = 0.95$, the ER design gave a similar probability of Type I error to that of the CA design. Consequently, for these designs, we used the same value of p_U. As expected, the CA design gave the lowest degree of covariate imbalance. Although we obtained a larger degree of covariate imbalance from the RACA design than from the CA design, it was considerably smaller than the values from the ER and RA designs.

Table 4.20: The values of $p_{x,k}$ and $\mu_{x,k}$ for all possible covariate combinations when k = 1, 2, 3, 4, in Scenario 3 when the parameters in Table 4.16 are used.

Arm	p_1	p_2	p_3	p_4	Z_1	Z_2	Z_3	μ_1	μ_2	μ_3	μ_4
A	0.2	0.4	0.1	0.3	0	0	0	5.21	44.70	99.48	134.29
					1	1	1	2.94	24.05	63.43	94.63
					1	0	0	4.26	36.97	85.63	115.58
					0	1	0	4.39	35.52	81.45	121.51
					0	0	1	4.26	36.60	90.02	121.51
					1	1	0	3.60	29.37	70.11	104.58
					0	1	1	3.60	29.08	73.70	109.95
					1	0	1	3.49	30.27	77.48	104.58
B	0.1	0.1	0.2	0.6	0	0	0	8.58	63.43	145.47	196.37
					1	1	1	4.85	34.12	92.76	138.38
					1	0	0	7.03	52.46	125.21	169.02
					0	1	0	7.24	50.40	119.10	177.68
					0	0	1	7.03	51.94	131.63	177.68
					1	1	0	5.93	41.68	102.51	152.93
					0	1	1	5.93	41.26	107.77	160.77
					1	0	1	5.75	42.95	113.30	152.93

Table 4.21: Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 (where the efficacies of arms A and B are equal) when using p_U = 0.95, N_{max} = 100, and n_0 = 20.

Design	Arm	P(selected)	α	Average (sd)	ANP	Degree of imbalance
ER	A B	0.016 0.020	0.020	50.055 (5.05) 49.945 (5.05)	100.00	0.200
RA	A B	0.070 0.069	0.069	47.031 (13.21) 46.805 (13.14)	93.84	0.210
RACA	A B	0.066 0.063	0.063	47.067 (13.60) 46.803 (13.68)	93.87	0.077
CA	A B	0.018 0.021	0.021	50.178 (7.27) 49.822 (7.27)	100.00	0.057

Table 4.22: Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 (where the efficacies of arms A and B are equal) when using $p_U = 0.95$, $N_{max} = 120$, and $n_0 = 30$.

Design	Arm	P(selected)	α	Average (sd)	ANP	Degree of imbalance
ER	A B	0.019 0.017	0.017	60.137 (5.45) 59.86 (5.45)	120.00	0.182
RA	A B	0.069 0.071	0.071	56.423 (14.81) 56.595 (14.96)	113.02	0.189
RACA	A B	0.069 0.063	0.063	56.606 (14.57) 56.927 (14.75)	113.53	0.062
CA	A B	0.017 0.018	0.018	60.157(7.67) 59.843 (7.67)	120.00	0.048

Table 4.23: Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 (where the efficacies of arms A and B are equal) when using $p_U = 0.95$, $N_{max} = 130$, and $n_0 = 30$.

Design	Arm	P(selected)	α	Average (sd)	ANP	Degree of imbalance
ER	A B	0.022 0.020	0.020	64.939 (5.74) 65.061 (5.74)	130.00	0.174
RA	A B	0.073 0.063	0.063	61.156 (16.26) 62.333 (16.39)	122.34	0.183
RACA	A B	0.066 0.072	0.072	61.185 (16.34) 61.040 (16.16)	122.23	0.061
CA	A B	0.022 0.021	0.021	64.956 (8.10) 65.044 (8.10)	130.00	0.044

Table 4.24: Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 when using different values of p_U, $N_{max} = 100$, and $n_0 = 20$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05.

Design	Arm	p_U	P(selected)	α	Average (sd)	ANP	Degree of imbalance
ER	A B	0.900	0.052 0.055	0.055	50.055 (5.05) 49.945 (5.05)	100.00	0.200
RA	A B	0.960	0.044 0.054	0.054	47.531 (12.95) 47.768 (12.83)	95.30	0.205
RACA	A B	0.955	0.054 0.057	0.057	47.464 (13.00) 47.669 (12.95)	95.13	0.072
CA	A B	0.900	0.053 0.057	0.057	50.178 (7.27) 49.822 (7.27)	100.00	0.057

Table 4.25: Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 when using different values of p_U, $N_{max} = 120$, and $n_0 = 30$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05.

Design	Arm	p_U	P(selected)	α	Average (sd)	ANP	Degree of imbalance
ER	A B	0.9000	0.057 0.052	0.052	60.137 (5.45) 59.863 (5.45)	120.00	0.182
RA	A B	0.9600	0.051 0.052	0.052	57.536 (14.01) 57.515 (14.21)	115.05	0.185
RACA	A B	0.9575	0.052 0.055	0.055	57.153 (14.02) 57.717(14.04)	114.87	0.060
CA	A B	0.9000	0.047 0.054	0.054	60.157 (7.67) 59.843 (7.67)	120.00	0.048

Table 4.26: Comparison of the simulation results for the ER, RA, RACA and CA designs, in Scenario 1 when using different values of p_U, $N_{max} = 130$, and $n_0 = 30$. The values of p_U have been selected to give a probability of Type I error of approximately 0.05.

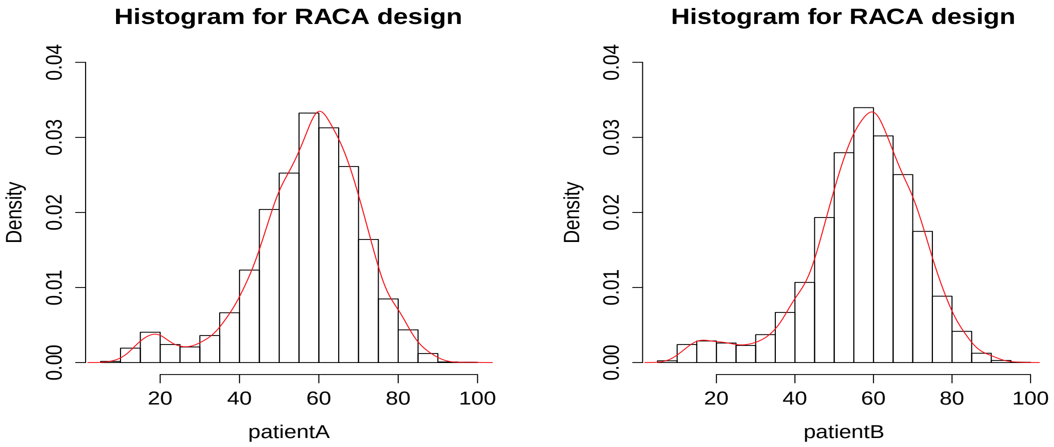
Design	Arm	p_U	P(selected)	α	Average (sd)	ANP	Degree of imbalance
ER	A B	0.905	0.054 0.054	0.054	64.939 (5.74) 65.061 (5.74)	130.00	0.174
RA	A B	0.960	0.051 0.052	0.052	62.452 (15.27) 62.333 (15.22)	124.78	0.178
RACA	A B	0.960	0.051 0.054	0.054	62.197 (15.23) 62.418 (15.36)	124.62	0.055
CA	A B	0.905	0.056 0.048	0.048	64.956 (8.10) 65.044 (8.10)	130.00	0.044

For the RACA and CA designs, a subsequent patient would be assigned to the treatment that would minimize the covariate imbalance. In contrast, for the ER and RA designs, we did not consider this factor when assigning a patient to a treatment. This may explain why the RACA and CA designs gave smaller covariate imbalances than the ER and RA designs.

Another advantage of the RACA design is that it requires a smaller ANP than the CA and ER designs. This benefit is similar to the property of the RA design. This is because, for the RACA and RA designs, the trials could stop early if one arm is demonstrably better than the other arm.

The results displayed in Tables 4.24 - 4.26 show that, for the RACA and RA designs, the average numbers of patients assigned to arms A and B were smaller than those for the CA design. In contrast, the standard deviations of the number of patients assigned to arms A and B were nearly twice as great as those for the CA design. We now provide some explanation of why this occurs.

To investigate the distribution of the number of patients assigned to arms A and B for the RACA and CA designs, the parameters shown in Table 4.25 were used to produce several histograms.

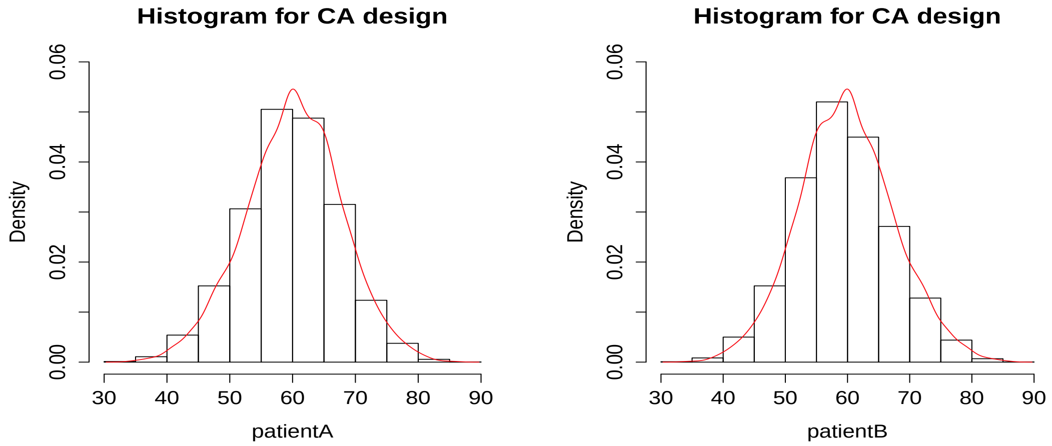


(a) the number of patients assigned to arm A (b) the number of patients assigned to arm B

Figure 4.1: The histograms of the number of patients assigned to arms A and B for the RACA design with $N_{max} = 120$.

The histograms shown in Figure 4.1 illustrate that, for the RACA design, there were two modes at 20 and 60 in the number of patients assigned to each arm. This is because, for the RACA design, there were two results for the trials. That is, some trials terminated early, while others stopped at the end of the study.

The first modes occurred at about 20 patients. However, the main modes occurred at about 60 patients. It can be seen that the primary modes were of considerably greater density than the secondary modes. In Scenario 1, the efficacies of the two arms are equal. Hence, there are small numbers of trials that show sufficient evidence that one arm is superior to another before the end of the trials. In contrast, the majority needs to progress until the end of the trials. The primary modes were at 60 because N_{max} was 120 and, as mentioned above, in Scenario 1 the efficacies of the two arms are equal. Consequently, about 60 patients were assigned to arm A and the others were assigned to arm B. For the reasons described above, it can be concluded from Figure 4.1 that, for the RACA



(a) the number of patients assigned to arm A (b) the number of patients assigned to arm B

Figure 4.2: The histograms of the number of patients assigned to arms A and B for the CA design with $N_{max} = 120$.

design, the average numbers of patients assigned to arms A and B were smaller than 60. There is also a considerable variation in the number of patients assigned to each arm.

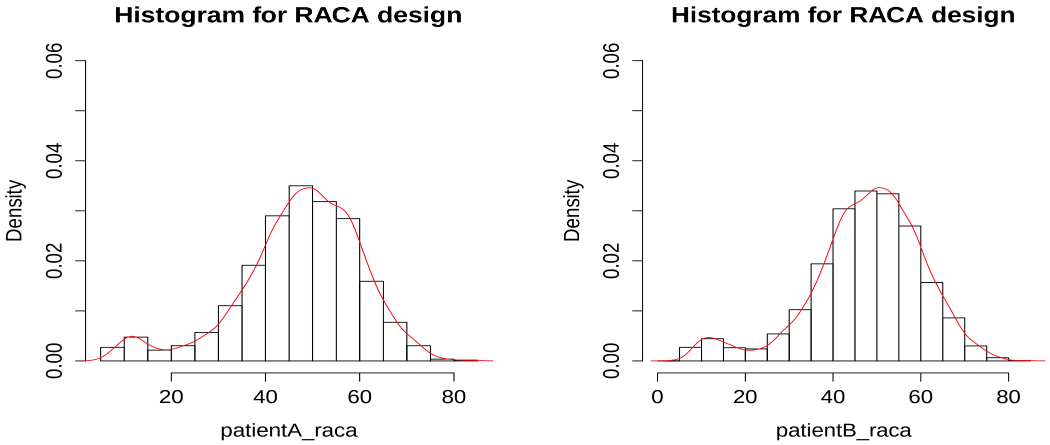
On the other hand, the histograms in Figure 4.2 show that, for the CA design, in both arms the histograms had just one mode at about 60. For the CA design, no trials can be terminated early so we have only one kind of trial. As a result, for the CA design, the average numbers of patients assigned to arms A and B were around 60. In addition, there was a slight variation in the numbers of patients assigned to each arm.

All the histograms above support the finding that, for the RACA design, the average numbers of patients assigned to arms A and B were smaller than those for the CA design. However, the standard deviations of the number of patients assigned to arms A and B were substantially greater than those for the CA design.

From Tables 4.24 - 4.26 again, the ANP increased as N_{max} was larger. On the other hand, the degree of imbalance decreased as N_{max} increased. For the ER and CA designs, the full quota of patients was used in the trial, and clearly, the

ANP would increase as N_{max} increased.

In order to examine the relationship between the ANP and N_{max} for the RACA and RA designs, histograms were produced of the numbers of patients assigned to arms A and B for the RACA design with N_{max} = 100, as shown in Table 4.24. These appear in Figure 4.3. Then we compared the histograms in Figure 4.1 with the histograms in Figure 4.3 (N_{max} = 100).



(a) the number of patients assigned to arm A (b) the number of patients assigned to arm B

Figure 4.3: The histograms of the number of patients assigned to arms A and B for the RACA design with N_{max} of 100.

The histograms in Figure 4.3 illustrate that, for the RACA design with N_{max} = 100, there are two modes at around 10 and 50 in the number of patients assigned to each arm. Hence, when N_{max} = 100, the two modes are less than those when N_{max} = 120. This may be a reason why the ANP increased as N_{max} increased for the RA and RACA designs.

When N_{max} is larger, we can better balance the covariates in the two treatments. Consequently, the degree of imbalance decreased as N_{max} increased.

Tables 4.27 - 4.29 show the results for the ER, RA, RACA and CA designs, in Scenario 2, when N_{max} = 100, 120 and 130 and n_0 = 20, 30 and 30 respectively. In addition, Tables 4.30-4.32 show the corresponding results in Scenario 3, when

$N_{max} = 100$, 120 and 130 and $n_0 = 20$, 30 and 30 respectively.

These results show that all designs gave similar powers of the test. We obtained the highest degree of imbalance from the RA design. Again, the CA design gave the lowest degree of imbalance. It is apparent that, for the RA and RACA designs in Scenarios 2 and 3, the degrees of covariate imbalance were higher than in Scenario 1. For the RA and RACA designs, a higher proportion of patients were allocated to a better treatment. In Scenario 1, the efficacies of arms A and B are equal, so the average numbers of patients assigned to arms A and B were equal. However, in Scenarios 2 and 3, arm B is better than arm A. Hence, in these scenarios the average numbers of patients going to arm A were substantially smaller than those going to arm B. This may be a reason why, for the RA and RACA designs in Scenarios 2 and 3, the degrees of imbalance were higher than in Scenario 1.

Once again when we look at Scenarios 2 and 3, we see that for the RACA and RA designs, the average numbers of patients assigned to arms A and B were less than for the ER and CA designs. We also see that, for the RACA and RA designs, the standard deviations of the number of patients assigned to arms A and B were higher than for the ER and CA designs.

The RA and RACA designs required substantially smaller ANP and ALT than the ER and CA designs. The AND was also less. This is because the first two designs can end early if one treatment is obviously better than the other. For the reason described earlier, the RA and RACA designs have higher PET than the ER and CA designs. They also have higher PBA than the ER and CA designs. As mentioned above, for the RA and RACA designs, one objective is to allocate more patients to the superior treatment. This may explain why the RACA and RA designs have higher PBA than the ER and CA designs.

As expected, for all designs, the power of the test, the ANP, the AND and the

ALT increased as N_{max} became larger. Also, the covariate imbalance decreased as N_{max} increased. For the RACA and RA designs, the PET rose as N_{max} increased. In contrast, for the CA and ER designs, the PET was 0 in all tables.

For the RACA and RA designs, the power of the test, the PET and the degree of covariate imbalance increased as the treatment effect increased. On the other hand, the ANP, the AND and the ALT decreased as the treatment effect increased.

For all designs, the PBA was consistent across the six tables. Although for the RACA and RA designs, one aim is to assign more patients to the better treatment, the PBA did not increase as the treatment effect increased. This is because in this chapter, we used $n_0 = 20$ and 30, whereas in the previous chapter we used n_0 = 1. Additionally, the ANP became smaller as the treatment effect increased. In particular, the difference between the ANP and n_0 became considerably smaller as the treatment effect increased.

It should be noted here that, for the RA design, the PBA is higher than that for the RACA design since, for the RA design, the assignment of a treatment to the next patient is based only on the response from the previous patients. However, for the RACA design, the decision to allocate a treatment to the next patient depends not only on the response from the previous patients but on the degree of covariate imbalance as well. This may explain why the PBA for the RA design is higher than that for the RACA design.

From Tables 4.27 - 4.32, it can be seen that the RACA design has power comparable to the other designs. It also uses the least ANP and ALT, and these are much less than those of the ER and CA designs. In addition, it has a degree of imbalance that is below that of the ER and RA designs. Moreover, the AND obtained from this design is considerably smaller than that from the ER and CA designs. The RACA design gave the highest PET. We see that, for the RACA design, more patients can be assigned to the superior treatment since the PBA is high, especially when compared with the ER and CA designs.

As far as these properties are concerned, the RACA and RA designs are slightly different. However, if the degree of covariate imbalance is of major concern, the RACA design is superior to the RA design.

4.5.1 Hypothesis testing

Using Tables 4.27 - 4.32, we conducted hypothesis testing as described in Section 2.2.1, to test whether these design characteristics and the covariate imbalance for the RACA design were different from those for the other designs.

For the degree of covariate imbalance, the decision that the difference in covariate imbalance is deemed to be practically significant will occur if the difference between the covariate imbalance obtained from the two designs is greater than 0.05. As all degrees of covariate imbalance were less than 1, 0.05 was selected as the minimum important difference for the degree of covariate imbalance.

It should be noted here that this value of 0.05 is less than the value of the minimum significant difference for the other mean design characteristics (the ANP, the AND and the ALT). For the other mean design characteristics, we regularly used four as the minimum important difference. Let

p_{er} denote a probability (e.g. the power) obtained from the ER design;
p_{ra} denote the corresponding probability obtained from the RA design;
p_{raca} denote the corresponding probability obtained from the RACA design;
p_{ca} denote the corresponding probability obtained from the CA design;
μ_{er} denote a mean (e.g the ANP) obtained from the ER design;
μ_{ra} denote the corresponding mean obtained from the RA design;
μ_{raca} denote the corresponding mean obtained from the RACA design;
μ_{ca} denote the corresponding mean obtained from the CA design;

Note that all results of hypothesis testing are based on 5,000 simulations.

The results of hypothesis testing when comparing the design characteristics and the degree of covariate imbalance obtained from the ER design and the RACA design are as follows:

- In the simulation whose results appear in Table 4.27, 95% CIs for the differences $p_{er} - p_{raca}$ for the power, for the PET and for the PBA were $(-0.007, 0.029)$, (-0.683, -0.657) and $(-0.129, -0.091)$ respectively. It could be seen that the CI for the difference in power lay entirely inside the interval $(-0.05, 0.05)$. The CIs for the differences in PET and PBA, however, lay completely outside the interval $(-0.05, 0.05)$. Similarly, for the simulation shown in Tables 4.28-4.32, the CIs for the difference in power (not shown here) lay absolutely inside the interval $(-0.05, 0.05)$, whereas the CIs for the differences in PET (not shown here) and PBA (not shown here), lay totally outside the interval $(-0.05, 0.05)$.

  It could be concluded that there was no difference between the powers, whether or not these powers were obtained from the ER design or the RACA design. In contrast, the PETs and the PBAs obtained from the two designs were different.

- Again in the simulation whose results appear in Table 4.27, 95% CIs for the differences $\mu_{er} - \mu_{raca}$ for the ANP, for the AND, and for the ALT were $(34.465, 36.215)$, $(6.064, 6.416)$ and $(60.723, 63.377)$ respectively. All CIs lay completely outside the interval $(-4, 4)$. For the simulation shown in Tables 4.28-4.32, we also obtained similar results to those from Table 4.27.

  We concluded that the ANPs, the ANDs, and the ALTs obtained from the two designs were different.

- In the simulation whose results appear in Table 4.27, a 95% CI for the difference $\mu_{er} - \mu_{raca}$ for the degree of covariate imbalance was $(0.032, 0.041)$. The CI for the covariate imbalance lay entirely inside the interval $(-0.05, 0.05)$. In addition, for the simulation shown in Tables 4.28-4.32, there were similar results to those from Table 4.27.

It could be concluded that there was no difference between the degrees of covariate imbalance, whether or not these covariate imbalances were obtained from the ER design or the RACA design.

When comparing the design characteristics and the degree of covariate imbalance obtained from the RA design and the RACA design, the results of hypothesis testing are as follows:

- In the simulation whose results appear in Table 4.27, 95% CIs for the differences $p_{ra} - p_{raca}$ for the power, for the PET and for the PBA were $(-0.054, -0.018)$, (-0.049, -0.011) and $(0.001, -0.091)$ respectively. Only the lower bound of the CI for the difference in power is less than - 0.05. The CIs for the differences in PET and PBA lay completely inside the interval $(-0.05, 0.05)$. For the simulation shown in Tables 4.28-4.32, there were similar results to those from Table 4.27.

  It could be concluded that there was no difference between the powers, the PETs and the PBAs obtained from the two designs.

- In the simulation whose results appear in Table 4.27, 95% CIs for the difference $\mu_{ra} - \mu_{raca}$ for the ANP, for the AND, and for the ALT were $(1.272, 3.747)$, $(-0.241, 0.161)$ and $(1.779, 5.541)$ respectively. The CIs for the differences in ANP, and AND lay completely inside the interval $(-4, 4)$. Only the upper bound of the CI for the difference in ALT is greater than 4. Additionally, for the simulation shown in Tables 4.28-4.32, there were similar results to that from Table 4.27.

We concluded that the ANPs, the ANDs, and the ALTs obtained from the two designs were similar.

- In the simulation whose results appear in Table 4.27, a 95% CI for the difference $\mu_{ra}-\mu_{raca}$ for the degree of covariate imbalance was $(0.096, 0.106)$. The CI for the difference in covariate imbalance lay entirely outside the interval $(-0.05, 0.05)$. For the simulation shown in Tables 4.28-4.32, we got similar results to those from Table 4.27.

  We concluded that the covariate imbalances obtained from the two designs were different.

Consider the results of hypothesis testing when comparing the design characteristics obtained from the CA design and the RACA design. We found that, in the simulation whose results appear in Table 4.27-4.32, 95% CIs for the differences $p_{ca} - p_{raca}$ for the power (not shown here) lay completely inside the interval $(-0.05, 0.05)$. On the other hand, the CIs for the differences in PET and PBA lay completely outside the interval $(-0.05, 0.05)$.

Again for the simulation shown in Table 4.27-4.32, 95% CIs for the differences $\mu_{ca}-\mu_{raca}$ for the ANP, for the AND, and for the ALT (not shown here) lay totally outside the interval $(-4, 4)$. Similarly, 95% CI for the difference $\mu_{ca} - \mu_{raca}$ for the degree of covariate imbalance lay entirely outside the interval $(-0.05, 0.05)$.

It could be concluded that the design characteristics and the degrees of covariate imbalance obtained from the CA design and RACA designs were different except in the power.

Table 4.27: Comparison of the results for the ER, RA, RACA and CA designs of Table 4.24, in Scenario 2 (arm B is better than arm A) when $N_{max} = 100$, and $n_0 = 20$.

Design	Arm	p_U	P(selected)	Power	Average (sd)	ANP	AND	PET	PBA	ALT	Degree of imbalance
ER	A	0.900	0	0.729	49.944 (5.01)	100.00	14.98	0	0.50	140	0.198
	B		0.729		50.056 (5.01)						
RA	A	0.960	0	0.682	24.739 (12.10)	67.17	8.83	0.64	0.63	81.61	0.263
	B		0.682		42.430 (21.37)						
RACA	A	0.955	0	0.718	25.496 (13.11)	64.66	8.74	0.67	0.61	77.95	0.162
	B		0.718		39.162 (20.81)						
CA	A	0.900	0	0.720	50.080 (7.35)	100.00	14.99	0	0.50	140	0.057
	B		0.720		49.920 (7.35)						

Table 4.28: Comparison of the results for the ER, RA, RACA and CA designs of Table 4.25, in Scenario 2 (arm B is better than arm A) when $N_{max} = 120$, and $n_0 = 30$.

Design	Arm	p_U	P(selected)	Power	Average (sd)	ANP	AND	PET	PBA	ALT	Degree of imbalance
ER	A	0.900	0	0.771	60.058 (5.52)	120.00	17.97	0	0.50	160	0.182
	B		0.771		59.942 (5.52)						
RA	A	0.960	0	0.735	29.857 (13.32)	78.65	10.54	0.69	0.62	91.04	0.240
	B		0.735		48.788 (24.24)						
RACA	A	0.9575	0	0.750	31.433 (14.49)	77.41	10.44	0.71	0.59	89.20	0.146
	B		0.750		45.973 (23.36)						
CA	A	0.900	0	0.769	60.100 (7.62)	120.00	17.98	0	0.50	160	0.048
	B		0.769		59.900 (7.62)						

Table 4.29: Comparison of the results for the ER, RA, RACA and CA designs of Table 4.26, in Scenario 2 (arm B is better than arm A) when $N_{max} = 130$, and $n_0 = 30$.

Design	Arm	p_U	P(selected)	Power	Average (sd)	ANP	AND	PET	PBA	ALT	Degree of imbalance
ER	A B	0.905	0 0.785	0.785	64.932 (5.73) 65.068 (5.73)	130.00	19.46	0	0.50	170.00	0.175
RA	A B	0.960	0 0.756	0.756	30.699 (14.54) 50.725 (26.79)	81.42	10.86	0.71	0.62	92.84	0.238
RACA	A B	0.960	0 0.762	0.762	32.087 (15.55) 48.210 (26.05)	80.30	10.80	0.73	0.60	91.17	0.144
CA	A B	0.905	0 0.776	0.776	65.092 (8.11) 64.908 (8.11)	130.00	19.48	0	0.50	170.00	0.044

Table 4.30: Comparison of the results for the ER, RA, RACA and CA designs of Table 4.24, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2) when $N_{max} = 100$,and $n_0 = 20$.

Design	Arm	p_U	P(selected)	Power	Average (sd)	ANP	AND	PET	PBA	ALT	Degree of imbalance
ER	A B	0.900	0 0.933	0.933	50.050 (5.31) 49.950 (5.31)	100.00	15.02	0	0.50	140.00	0.197
RA	A B	0.960	0 0.902	0.902	21.018 (10.08) 37.260 (19.51)	58.28	7.39	0.84	0.64	64.58	0.279
RACA	A B	0.955	0 0.911	0.911	22.148 (11.39) 33.758 (18.64)	55.91	7.22	0.86	0.60	61.72	0.180
CA	A B	0.900	0 0.932	0.932	50.094 (7.09) 49.906 (7.09)	100.00	14.93	0	0.50	140	0.057

Table 4.31: Comparison of the results for the ER, RA, RACA and CA designs of Table 4.25, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2) when $N_{max} = 120$, $n_0 = 30$.

Design	Arm	p_U	P(selected)	Power	Average (sd)	ANP	AND	PET	PBA	ALT	Degree of imbalance
ER	A B	0.900	0 0.966	0.966	59.963 (5.42) 60.037 (5.42)	120.00	17.99	0	0.50	160.00	0.180
RA	A B	0.960	0 0.944	0.944	24.458 (10.37) 39.677 (20.97)	64.14	8.42	0.90	0.62	68.03	0.263
RACA	A B	0.9575	0 0.946	0.946	26.000 (11.80) 37.473 (19.96)	63.58	8.30	0.91	0.59	67.35	0.176
CA	A B	0.900	0 0.964	0.964	60.100 (7.62) 59.900 (7.62)	120.00	17.96	0	0.50	160	0.048

Table 4.32: Comparison of the results for the ER, RA, RACA and CA designs of Table 4.26, in Scenario 3 (arm B is better than arm A and the treatment effects are higher than in scenario 2) when $N_{max} = 130$, $n_0 = 30$.

Design	Arm	p_U	P(selected)	Power	Average (sd)	ANP	AND	PET	PBA	ALT	Degree of imbalance
ER	A B	0.905	0 0.970	0.970	65.044 (5.69) 64.956 (5.69)	130.00	19.50	0	0.50	170.00	0.174
RA	A B	0.960	0 0.954	0.954	25.010 (10.94) 40.996 (22.629)	66.01	8.54	0.92	0.62	69.19	0.260
RACA	A B	0.960	0 0.958	0.958	26.594 (12.41) 39.020 (21.38)	65.61	8.45	0.92	0.59	68.66	0.172
CA	A B	0.905	0 0.968	0.968	64.813 (8.05) 65.19 (8.05)	130.00	19.49	0	0.50	170	0.044

4.5.2 A comparison of expected costs

Refer to Section 3.4 for a description of expected cost. From Table 4.24, in the event that H_0 is true, the expected cost (3.1) mentioned in Section 3.4 can be written as follows:

the ER design

$$E(\text{cost}) = 100.00K_2 + 50.055K_a + 49.945K_b + 0.055K_0,$$

the RA design

$$E(\text{cost}) = 95.30K_2 + 47.531K_a + 47.768K_b + 0.054K_0,$$

the RACA design

$$E(\text{cost}) = 95.13K_2 + 47.464K_a + 47.669K_b + 0.057K_0,$$

the CA design

$$E(\text{cost}) = 100.00K_2 + 50.178K_a + 49.822K_b + 0.057K_0.$$

The equations from the other tables follow similarly. For this example, it can be seen that, if H_0 was true, the ER cost exceeds the RA cost. Similarly, the CA cost exceeds the RACA cost for positive K_0, K_2, K_a and K_b.

In other examples, where we obtain different coefficient of K_0, this might not be true.

From Table 4.27, in the event that H_0 is false, the expected cost (3.2) mentioned in section 3.4 can be written as follows:

the ER design

$$E(\text{cost}) = 100.00K_2 + 49.944K_a + 50.056K_b + 0.271K_1,$$

the RA design

$$E(\text{cost}) = 67.17K_2 + 24.739K_a + 42.430K_b + 0.318K_1,$$

the RACA design

$$E(\text{cost}) = 64.66K_2 + 25.496K_a + 39.162K_b + 0.282K_1,$$

the CA design

$$E(\text{cost}) = 100.00K_2 + 50.080K_a + 49.920K_b + 0.280K_1.$$

The equations from the other tables follow similarly. As mentioned in Section 3.4, if H_0 is false, there are two principal parts of the expected cost: the constant cost and the cost of lost opportunity from accepting H_0 if H_0 is false. Using the RACA and RA designs can reduce the constant cost substantially compared to the ER and CA designs. On the other hand, they have a higher cost of lost opportunity from accepting H_0 if H_0 is false than the other two designs. If the value of K_1 is considerably larger than the values of K_2, K_a and K_b, using the ER and CA designs has a lower cost than using the adaptive design. Otherwise, the RACA and RA designs have lower costs than the other two designs.

4.5.3 A comparison of expected total costs

The results so far show that although the RACA and CA designs give lower power of test than the ER and CA designs, the difference is minor. Moreover, for economical and ethical reasons, the RACA and RA designs have an advantage over the ER and CA designs. Consequently, in this section, we intend to compare the expected total costs of the RACA and RA designs.

From equation (3.8), the expected total cost of the RACA design is

$$E(\text{total cost}) = A + BP_{H_0}; \tag{4.13}$$

where $A = 2n_{2RACA}K_2 + n_{a2RACA}K_a + n_{b2RACA}K_b + \beta_{RACA}K_1$ and $B = 2K_2(n_{1RACA} - n_{2RACA}) + \alpha_{RACA}K_0 - \beta_{RACA}K_1 + K_a(n_{a1RACA} - n_{a2RACA}) + K_b(n_{b1RACA} - n_{b2RACA})$.

Similarly, the expected total cost of the RA design is

$$E(\text{total cost}) = C + DP_{H_0}; \tag{4.14}$$

where $C = 2n_{2RA}K_2 + n_{a2RA}K_a + n_{b2RA}K_b + \beta_{RA}K_1$ and $D = 2K_2(n_{1RA} - n_{2RA}) + \alpha_{RA}K_0 - \beta_{RA}K_1 + K_a(n_{a1RA} - n_{a2RA}) + K_b(n_{b1RA} - n_{b2RA})$.

We use the same example as in Section 3.8. That is, suppose $K_0 = 490,000$, $K_1 = 440,000$, $K_2 = 50$ and $K_a = K_b = 100$.

From the results in Table 4.26 and Table 4.32 in Section 4.5, firstly, A, B, C and D are calculated.

$$\begin{aligned}
A &= (2 \times 65.61 \times 50) + (26.59 \times 100) + (39.02 \times 100) + (0.04 \times 440,000) \\
&= 30,722 \\
B &= 100 \times (124.62 - 65.61) + (0.05 \times 490,000) - (0.04 \times 440,000) \\
&\quad +100 \times (62.2 - 26.59) + 100 \times (62.42 - 39.02) \\
&= 18,702 \\
C &= (2 \times 66.01 \times 50) + (25.01 \times 100) + (41 \times 100) + (0.05 \times 440,000) \\
&= 35,202 \\
D &= 100 \times (124.78 - 66.01) + (0.05 \times 490,000) - (0.05 \times 440,000) \\
&\quad +100 \times (62.45 - 25.01) + 100 \times (62.33 - 41) \\
&= 14,254
\end{aligned}$$

Since $B \neq D$ in this case, then the value of $x = P_{H_0}$ at the point of intersection is calculated.

$$\begin{aligned} P_{H_0} &= \frac{C-A}{B-D} \\ &= 1.01 \end{aligned}$$

In this example, $x = \frac{C-A}{B-D}$ does not lie between 0 and 1. Hence, the RA design has greater expected total cost than the RACA design for all values of P_{H_0}. Figure 4.4 shows an illustrative example of the expected total costs of the RACA and RA designs. It can be seen that the expected total costs of the RACA design is less than that of the RA design for all values of P_{H_0}. This is because the RACA gives higher power than the RA design. Additionally, in this example, the value of K_1 is considerable higher than those of K_2, K_a and K_b, so A and B are lower than C and D.

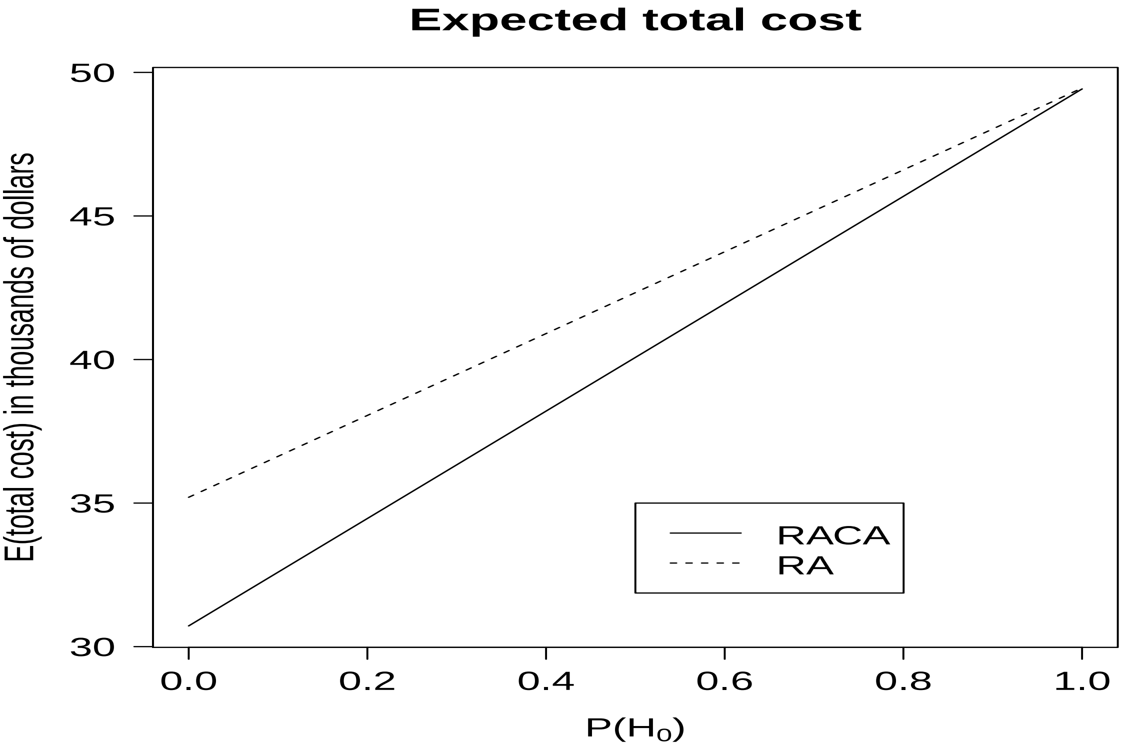


Figure 4.4: E(total cost) of the RACA and RA designs as a function of P_{H_0}.

From this example, it is suggested that the RACA design is better than the RA design because of the lower expected total cost and the lower degree of imbalance.

4.6 Conclusion

It can be clearly seen from statistical, ethical and economic perspectives that the RACA design is the best design to use, since this design gives a power of the test that is similar to those of the other designs. Moreover, the resources required by this design are less than those required by the other three designs. Additionally, in this design, the average number of patients allocated to an inferior treatment is small. Even though the degrees of imbalance obtained from this design are higher than those of the CA design, they are not worse than those of the ER and RA designs.

Chapter 5

Conclusions

5.1 Summary

Clinical trials are research studies which involve both healthy people and patients. The aim of these trials is to evaluate the efficacy of a new treatment or to compare the efficacy of new treatments with the existing treatment. In this thesis, our main concern is to compare the efficacy of a new treatment with a current standard treatment.

An adaptive design is a clinical trial design that allows changes to a trial by using accrued data. This data will be used to make a decision on how to change a trial without affecting its validity and integrity. Adaptive designs have been used throughout this thesis. We focused on adaptive randomisation and interim analyses. In particular, we focused on response-adaptive randomisation and covariate adaptive randomisation.

One aim of this thesis was to extend and generalise the adaptive designs of Huang *et al.* (2009) (HNL). We began by examining the two aspects (1) the enrolment regime and (2) the randomisation procedure, by considering the response-adaptive randomisation and the degree of covariate imbalance. We also intended to use a generalised linear model to introduce a scheme to the HNL design which enables covariates to be considered. In this thesis, three covariates, Z_1, Z_2 and Z_3, were considered. The distributions of Z_2 and Z_3 were conditional on the value of Z_1. In real life, it is difficult to find independent covariates. Some covariates may depend on others.

Furthermore, in this thesis, the covariates were considered in conjunction with one another. That is, the determination of imbalance of pairs of covariates was performed. In reality, the combination of characteristics of patients jointly affects a response variable.

Additionally, we addressed important criteria for evaluating and comparing a design with competing designs. Then an application for using these criteria for assessing and comparing the designs was provided.

Chapter 2 examined the adaptive method of HNL. One assessment was performed by considering a different recruitment regime for this method. In this enrolment regime, the accrual rate was changed from exactly one patient per week to an average of one per week. In reality, it is rare to find that patients come into the trial at a rate of exactly one per week. Thus an investigation into whether this more realistic scenario affects the results obtained from simulation was carried out. It was found that the differences between the statistical properties of the two enrolment regimes (i.e. exactly one new patient per week, and an average of one new patient per week) should not be considered practically significant. We concluded that the HNL practice is a sensible approach to use, and continued to follow their practice of having exactly one arrival per week.

Chapter 3 addressed principal criteria for evaluating and comparing designs. We focused on several criteria: the Operating Characteristic Curve, and various design characteristics. We found that the OC curve is not an appropriate method of comparing clinical trial designs due to an additional complication. That is that the OC curve is not uniquely defined by $\mu_a - \mu_b$.

Eight design characteristics were considered in this chapter. We focused on the power of the test, PBA, PET, ANP, AND, ALT, expectation cost and the expected total costs of the designs. In order to obtain design characteristics, two situations were considered: we have no prior knowledge of how arms A and B differ, and we suppose that treatment B is superior to treatment A. The hypotheses were

$\mathrm{H}_0 : \mu_a \geq \mu_b$;

$\mathrm{H}_1 : \mu_a < \mu_b$.

By using the eight design characteristics, the major finding of this chapter was that an adaptive design that uses the response-adaptive randomisation is better than the equal randomization (ER) design. This is because, for economic and ethical reasons, the adaptive design has an advantage over the ER design. The economic reason is that the adaptive design requires considerably fewer resources than the ER design. The ethical reason is that the adaptive design uses a smaller ANP than the ER design. Additionally, we can obtain a smaller AND and a larger PBA from the adaptive design than from those of the ER design. Hence, using the adaptive design can reduce the number of patients who are involved in the trial and receiving an inferior treatment. Although the adaptive design gave lower power of the test than the ER, the difference was minor. The lower power of the test obtained from the adaptive design can be traded off against the two advantages of the adaptive design described above.

In conclusion, as far as economic and ethical reasons are concerned, the adaptive design is better than the ER design. If the main concern is statistical power, the adaptive design is a competitive design. Overall, we have confirmed that the adaptive design is a better design when trading off between the two reasons and statistical power.

Chapter 4 extended the HNL design to an applicable design. That is, we enabled it to work in a more realistic situation. The following improvements were made:

- we developed an appropriate randomisation procedure by considering the response of the previous patients and the degree of covariate imbalance.
- we ensured that a subsequent patient will be more likely not only to receive better treatment but also to minimize the degree of covariate imbalance.

By using this procedure, we can minimize bias from covariate imbalance and provide more efficient comparison of treatment effect.

In order to enable covariates to be considered in the HNL design, a generalised linear model was used. In this research, our response variable is $T^{k}_{x,i}$ which is the progression-free survival time of participant i in arm x if this patient occupies the kth category. Recall that $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$ is the mean progression-free survival time of the kth category in arm x. We had $T^{(k)}_{x,i} \sim \text{Exp}(\lambda_{x,k})$ and $\text{E}(T^{(k)}_{x,i}) = \frac{1}{\lambda_{x,k}}$. Therefore, $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$ can be fitted as a model via the log link by using a generalised linear model.

In this research, the values of $\mu_{x,k}$ was generated by $\mu_{x,k} = \frac{1}{\lambda_{x,k}} = \exp(\beta_{0k} + \beta_{Tk}T + \beta_{1k}Z_1 + \beta_{2k}Z_2 + \beta_{3k}Z_3)$ where T is an indicator variable; $T = 0, 1$ if patient is allocated to treatment A or B, the quantity k is a category variable: $k = 1, 2, 3, 4$; Z_1, Z_2 and Z_3 are three binary covariates; β_{0k} is the intercept of the kth category, β_{Tk} is the treatment coefficient of the kth category, and β_{1k}, β_{2k}, β_{3k} are the coefficient of Z_1, Z_2 and Z_3 in the kth category .

Since we aimed to extend the HNL design to perform in a more realistic situation, we proceeded as follows. The distributions of Z_2 and Z_3 were conditional on the value of Z_1. In real life, for example, the number of previous chemotherapy treatments may depend upon the age of the patients. Moreover, we considered the covariates in conjunction with one another. In reality, the combination of characteristics of patients affects a response variable simultaneously. For instance, gender and smoking may jointly affect hypertension. Therefore, in this thesis, the degree of imbalance of covariates was determined in pairs.

In Chapter 3, we found which an adaptive design that uses the response-adaptive randomisation (RA) is better than the ER design. Hence, in Chapter 4, we compared the RA with an adaptive design (RACA) that was proposed in this chapter by employing the design characteristics mentioned in Chapter 3 and the degree of covariate imbalance.

As far as these characteristics are concerned, the performances of RACA and RA designs are slightly different. However, if the degree of covariate imbalance is of principal concern, the RACA design is superior to the RA design. In conclusion, the RACA design is the best design.

Overall, this thesis has increased the understanding of the properties of adaptive designs by including principal criteria for evaluating and comparing designs. Additionally, the extended design takes better account of covariates. It is worthwhile to consider a randomisation procedure that combines the response of the previous patients with the degree of covariate imbalance because this design can decrease bias and provide more effective comparisons. Moreover, the proposed design is more realistic, because it considers covariates simultaneously and allows them to be dependent upon one another.

5.2 Discussion

Although the RACA designs and the CARA designs (Hu and Rosenberger, 2006) have similar abbreviations, the procedures for randomizations in the two designs are different. In the RACA designs, the probability of assigning a treatment to a current patient is based on both the response and the degree of covariate imbalance of the previous patients. In the CARA designs, the current patient is allocated to a treatment by considering the history of previous patients' treatment assignments (responses and covariates) as well as the values of covariates of the current patient. The important difference between the RACA and the

CARA designs is that in the CARA designs, the logistic regression model was used to determine the probability of assigning a new patient to the treatment. This model was given by including the treatment-covariate interactions term. The probability of assigning a new patient to treatment A was based on the estimated covariate odds ratio. However, in this thesis, a generalised linear model is utilized for enabling covariates to be considered in the HNL designs. We did not consider the treatment-covariate interactions term in the generalised linear model. The probability of assigning a new patient to treatment A was based on both the posterior probability evaluated while the trial progressed, and the degree of covariate imbalance.

In this thesis, we generalised extensively the adaptive designs of HNL. The developed design is applicable to more realistic situations.

In the HNL designs, the response adaptive randomisation is used. That is, the probability of allocating a new patient to a treatment is based only upon the response of the previous patients. However, this randomisation method does not consider the possibility that important prognostic factors might influence the effect of the treatments. This omission may cause bias. In order to minimize bias and provide an efficient comparison of the treatments, the randomization procedure was improved. The developed randomization procedure can increase the benefits of the HNL design because a subsequent patient will be more likely to receive the better treatment. The design will also minimize the degree of covariate imbalance.

Due to this, we used a generalised linear model to enable covariates to be considered in the HNL designs.

Conventionally, in designs, the covariates are assumed to be independent of one another. This is because it is simpler to produce a design under this assumption. However, it is rare to find such a situation in reality. In general, some covariate may depend on another. For instance, older people may have a higher probability of having hypertension than younger people. Hence it is valuable to consider the practical situation in which some covariates depend on others even though it adds complications to the design process. In this thesis, the conditional probability is utilized for generating the dependent covariates.

We developed a new approach for determining the degree of covariate imbalance. In HNL, the degree of covariate imbalance was determined separately for each covariate. In reality, a response variable is affected by a combination of the characteristics of patients simultaneously. Due to this, we considered the covariates in conjunction with one another. Hence, in the new approach, the measurement of imbalance was carried out in pairs.

The two paragraphs given above, demonstrate that in this thesis, more realistic situations were considered when producing a design.

In addition, a more realistic enrolment regime was investigated. The arrival rate was changed from exactly one patient per week to an average of one per week. In reality, an accrual schedule of exactly one patient per week rarely happens. A simulation study was then carried out to investigate whether this more realistic scenario affects the results. It was found that the more realistic scenario has no significant effect on the outcome, thereby providing a justification previously not given for using the exact arrival pattern.

When designing a clinical trial, one important step is to evaluate and compare a proposed design with other designs. By doing this, researchers can ensure that the proposed design is effective. Hence, some important criteria for evaluating and comparing designs were investigated and employed.

In this thesis, although for two competing designs the difference in mean $(\mu_A - \mu_B)$ was identical, the power of the test was found to be not automatically the same. Hence, the power is not a function of $(\mu_A - \mu_B)$. A consideration of the difference $(\mu_A - \mu_B)$ is thus not sufficient to compare the power of competing designs of the kind considered in this thesis.

It was found that from statistical, ethical and economic perspectives the RACA design is the best design to use, since this design gives a power of the test that is similar to those of the other designs. Additionally, the resources required by this design are less than those required by the ER, RA and CA designs. Furthermore, in the RACA design, the average number of patients allocated to an inferior treatment is small. Although the degrees of covariate imbalance obtained from this design are higher than those of the CA design, they are not worse than those of the ER and RA designs.

Hence, the RACA designs can combine features of a good design covering efficiency and ethics and include balance of covariates.

In conclusion, it can be seen that our thesis is valuable. This is because by statistical, ethical and economic perspectives, our proposed design is a good design and is applicable to more realistic situations.

5.3 Future research

1. In this thesis, we mentioned two possible ways to enable covariates to be considered in the RA design. We can fit models for $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$, or for $p_{x,k}$ where $p_{x,k}$ is the probability of a patient in arm x occupying the kth category of a short-term response. However, in this thesis, we only focused on fitting a model for $\mu_{x,k} \equiv \frac{1}{\lambda_{x,k}}$. Hence, in future research, $p_{x,k}$ could be fitted and the different models affect the results obtained when simulation is carried out could be investigated.

   Since $(S_{x,1,i}, ..., S_{x,4,i}) \sim \text{Multi}\ (1, p_{x,1}, ..., p_{x,4})$ for $k = 1, 2, 3, 4$ and $\mathrm{E}(S_{x,k}) = p_{x,k}$, we can fit a model by using the Multinomial logit model. It was suggested by Faraway (2006) that since $\sum_{k=1}^{4} p_{x,k} = 1$, one category should be chosen as a baseline. Hence, $p_{x,1}$ is given by

$$p_{x,1} = 1 - \sum_{k=2}^{4} p_{x,k}.$$

The link function used is the logit which is $\eta_k = \ln(\frac{p_{x,k}}{p_{x,1}}), k = 2, 3, 4$. Faraway (2006) suggested that by using this function, we can ensure that $0 \leq p_{x,k} \leq 1$.

Therefore, $p_{x,k}$ is based on a model which is given by

$$\ln(\frac{p_{x,k}}{p_{x,1}}) = \beta_{0k} + \beta_{Tk}T + \beta_{1k}Z_1 + \beta_{2k}Z_2 + \beta_{3k}Z_3, \tag{5.1}$$

where the variables on the right-hand side are as defined on page 154.

The model (5.1) can be rewritten as

$$\frac{p_{x,k}}{p_{x,1}} = \exp(\eta_k) = \exp(\beta_{0k} + \beta_{Tk}T + \beta_{1k}Z_1 + \beta_{2k}Z_2 + \beta_{3k}Z_3),$$

It follows that

$$p_{x,k} = \frac{\exp(\eta_k)}{1 + \sum_{k=2}^{4} \exp(\eta_k),} k = 2, 3, 4$$

and

$$p_{x,1} = \frac{1}{1 + \sum_{k=2}^{4} \exp(\eta_k)}.$$

This could be generalised to more than four categories.

2. Originally, in the CA procedure, Ning and Huang (2010) determined the degree of imbalance of each covariate separately. Their program only allowed for determining one covariate. On the other hand, we considered the degree of imbalance of covariates in pairs. Hence, we needed to write a new program which allowed the examination of pairs of covariates. It will be useful if we have a general program that can be used whatever the number

of covariates is. We then do not need to write a new program every time the number of covariates is changed. Consequently, further research for this program is required.

3. In this thesis, the method of the degree of covariates imbalance used is for discrete covariates. However, two of the covariates considered (patient age and the number of previous chemotherapy treatments) are continuous. Thus, we needed to divide them into two categories. It might be more appropriate if we can balance the continuous covariates directly.

   Ciolio *et al.* (2011) considered some aspects of balancing continuous covariates. However, they did not considered more than two covariates and they always assumed that these covariates were independent. This is unlikely to be a realistic assumption. In future research, continuous covariates should be balanced directly.

4. In this thesis, after each patient was enrolled in the trial, we assumed that he or she was followed until the end of the trial or the trial was terminated. In real life, however, it always happens that some patients may drop out of the trial. Hence, this situation might be considered and simulated in future research.

Appendix A

Main R program used

This program was obtained by modifying Ning (2009) and the Ning2.

```
#set values of parameters pfavour 0.8
#senario2

ininum <- 30 #number of patients randomly allocated to treatments
ntotal <- 120 #total number of patients
nsimul <- 5000#total number of simulations of all patients
p1 <- 0.7 #P(X1=1)
p20 <- 0.4 #P(X2=1|X1=0)
p21 <- 0.65 #P(X2=1|X1=1)
p30 <- 0.5 #P(X3=1|X1=0)
p31 <- 0.6 #P(X3=1|X1=1)
pfavour <- 0.8

gamma<-c(0.5,0.5,0.5,0.5)
rectime<-rep(1:ntotal)
alpha<-c(11,11,11,11)
beta<-c(40,300,750,1100)

addtime<-40

#k=1
b01<-1.65
b11<--0.2
b21<--0.17
b31<--0.2
bt1<-0

#k=2
b02<-3.8
b12<--0.19
```

```
b22<--0.23
b32<--0.2
bt2<-0

#k=3
b03<-4.6
b13<--0.15
b23<--0.2
b33<--0.1
bt3<-0

#k=4
b04<-4.9
b14<--0.15
b24<--0.1
b34<--0.1
bt4<-0
catnum<-4

underp1<-c(0.2, 0.4, 0.1, 0.3)
underp2<-c(0.1, 0.1, 0.2, 0.6)

Pu<-0.9575
#Pl<-0.1

#imbalance <- 0
library(lattice)
library(MASS)
library(mvtnorm)
library(coda)

library(pscl)
library(MCMCpack)
###########################################################################

comp<-function(a0,a1,b0,b1,pp1,pp2)
{

  ppp1<-rdirichlet(10000,pp1)
  ppp2<-rdirichlet(10000,pp2)

  tempx<-rep(0,10000)
  tempy<-rep(0,10000)
  for(k in 1:length(a0))
  {
    tempx<-tempx+ppp1[,k]*rigamma(10000,a0[k],b0[k])
    tempy<-tempy+ppp2[,k]*rigamma(10000,a1[k],b1[k])}

  result<-mean(ifelse(tempx<tempy,1,0))

  return(result)

}
##########################################################################

fchose<-rep(0, nsimul)# treat is chosen as superior in the final simulation
chose<-rep(0, nsimul)# treat is chosen as superior before the final simulatio
treatpro<-rep(0, nsimul)
nlength<-rep(0, nsimul)
imbalance<-rep(0,nsimul)
patientA_raca<-rep(0,nsimul)
patientB_raca<-rep(0,nsimul)


duration<-rep(0,nsimul)
early<-rep(0,nsimul)
death<-rep(0,nsimul)
maxtime<-rectime[length(rectime)]+addtime
for (isim in 1:nsimul)
{
  #generate covariates
  X1 <- rbinom(ntotal,1,p1)
  X2 <- rep(0,ntotal)
  X3 <- rep(0,ntotal)
  treat <-NULL #rep(0,ininum)
  indic <- (X1==1)
  nX11 <- sum(indic)
  nX10 <- ntotal -  nX11
  X2[indic] <- rbinom(  nX11,1,p21)
  X2[!indic] <- rbinom(nX10,1,p20)
  X3[indic] <- rbinom(  nX11,1,p31)
  X3[!indic] <- rbinom(nX10,1,p30)

  #table12 <- table(X1,X2)
  #table13 <- table(X1,X3)
  #prop.table(table12,1)
  #prop.table(table13,1)
  treatshort <- rbinom(ininum,1,0.5)

  status1<-rmultinom(ininum, size = 1, prob=underp1)
```

```
status2<-rmultinom(ininum, size = 1, prob=underp2)

st1<-rep(0,ininum)
st2<-rep(0,ininum)

for(kkk in 1:catnum)
{
  st1<-st1+status1[kkk,]*kkk
  st2<-st2+status2[kkk,]*kkk
}
st<-treatshort*st2+(1-treatshort)*st1

log<-rep(0,ininum)
elamda<-rep(0,ininum)
T<-rep(0,ininum)
for(jjj in 1:ininum)
{
  if(st[jjj]==1){
    log[jjj]<-b01+bt1*treatshort[jjj]+b11*X1[jjj]+b21*X2[jjj]+b31*X3[jjj]
    elamda[jjj]<-1/exp(log[jjj])
  } else if (st[jjj]==2){
    log[jjj]<-b02+bt2*treatshort[jjj]+b12*X1[jjj]+b22*X2[jjj]+b32*X3[jjj]
    elamda[jjj]<-1/exp(log[jjj])
  } else if (st[jjj]==3){
    log[jjj]<-b03+bt3*treatshort[jjj]+b13*X1[jjj]+b23*X2[jjj]+b33*X3[jjj]
    elamda[jjj]<-1/exp(log[jjj])
  } else if (st[jjj]==4){
    log[jjj]<-b04+bt4*treatshort[jjj]+b14*X1[jjj]+b24*X2[jjj]+b34*X3[jjj]
    elamda[jjj]<-1/exp(log[jjj])
  }
  T[jjj]<-rexp(1,rate=elamda[jjj])

}#loop of generating lamda

TT<-cbind(T,rectime[1:ininum]+T)


treat <- treatshort

#construct the initial incidence matrices after ininum patients
treat1 <- (treatshort==1)
X1short <- X1[1:ininum]
X1short0 <- X1short[!treat1]
X1short1 <- X1short[treat1]
X2short <- X2[1:ininum]
```

```
X2short0 <- X2short[!treat1]
X2short1 <- X2short[treat1]
X3short <- X3[1:ininum]
X3short0 <- X3short[!treat1]
X3short1 <- X3short[treat1]
mat012 <- table(factor(X1short0, levels=c(0,1)),factor(X2short0,
          levels=c(0,1)))
mat013 <- table(factor(X1short0, levels=c(0,1)),factor(X3short0,
          levels=c(0,1)))
mat023 <- table(factor(X2short0, levels=c(0,1)),factor(X3short0,
          levels=c(0,1)))
mat112 <- table(factor(X1short1, levels=c(0,1)),factor(X2short1,
          levels=c(0,1)))
mat113 <- table(factor(X1short1, levels=c(0,1)),factor(X3short1,
          levels=c(0,1)))
mat123 <- table(factor(X2short1, levels=c(0,1)),factor(X3short1,
          levels=c(0,1)))


stop<-0
j<-ininum+1
upp1<-rep(0,catnum)
upp2<-rep(0,catnum)
t0<-rep(0,catnum)
t1<-rep(0,catnum)
n0<-rep(0,catnum)
n1<-rep(0,catnum)
while  (j <= ntotal & stop == 0)
{
  TTT<-ifelse(TT[,2]<=rectime[j], TT[,1], (rectime[j]+TT[,1]-TT[,2]))

  for(cc in 1:catnum)
  {
    upp1[cc]<-gamma[cc]+sum(ifelse(treat==0 & st==cc ,1,0))
    upp2[cc]<-gamma[cc]+sum(ifelse(treat==1 & st==cc ,1,0))
    t0[cc]<-sum(TTT[treat==0 & st==cc])
    t1[cc]<-sum(TTT[treat==1 & st==cc])
    n0[cc]<-length(TT[(TT[,2]<=rectime[j] & treat==0 & st==cc),1])
    n1[cc]<-length(TT[(TT[,2]<=rectime[j] & treat==1 & st==cc),1])
  }

  d<-n0[1]+n1[1]

  ga1<-alpha+n0
  ga2<-alpha+n1
```

```
gb1<-beta+t0
gb2<-beta+t1
probtemp<-comp(ga1,ga2,gb1,gb2,upp1,upp2)
probra<-sqrt(probtemp)/(sqrt(probtemp)+sqrt(1-probtemp))


#What happens if we make the next treatment 0?
newmat012 <- mat012
newmat013 <- mat013
newmat023 <- mat023
newmat012[1+X1[j],1+X2[j]] <- newmat012[1+X1[j],1+X2[j]] + 1

newmat013[1+X1[j],1+X3[j]] <- newmat013[1+X1[j],1+X3[j]] + 1
newmat023[1+X2[j],1+X3[j]] <- newmat023[1+X2[j],1+X3[j]] + 1
#Table 4.12 page115
M12 <- matrix(as.vector(cbind(t(newmat012),t(mat112))),4,2)
M13 <- matrix(as.vector(cbind(t(newmat013),t(mat113))),4,2)
M23 <- matrix(as.vector(cbind(t(newmat023),t(mat123))),4,2)
#E12, E13 and E23 are the matrices of expected
#values of M12, M13 and M23 respectively
E12 <- outer(rowSums(M12),colSums(M12)[2],FUN = "*")/j
E13 <- outer(rowSums(M13),colSums(M13)[2],FUN = "*")/j
E23 <- outer(rowSums(M23),colSums(M23)[2],FUN = "*")/j
DA <- sum(abs(M12[,2]-E12), abs(M13[,2]-E13), abs(M23[,2]-E23))/j
# cat("  E12 :",  E12 ,"\n")
#cat("  E13 :",  E13 ,"\n")
#cat("  E23 :",  E23 ,"\n")
#What happens if we make the next treatment 1?
newmat112 <- mat112
newmat113 <- mat113
newmat123 <- mat123
newmat112[1+X1[j],1+X2[j]] <- newmat112[1+X1[j],1+X2[j]] + 1
newmat113[1+X1[j],1+X3[j]] <- newmat113[1+X1[j],1+X3[j]] + 1
newmat123[1+X2[j],1+X3[j]] <- newmat123[1+X2[j],1+X3[j]] + 1
M12 <- matrix(as.vector(cbind(t(mat012),t(newmat112))),4,2)
M13 <- matrix(as.vector(cbind(t(mat013),t(newmat113))),4,2)
M23 <- matrix(as.vector(cbind(t(mat023),t(newmat123))),4,2)
E12 <- outer(rowSums(M12),colSums(M12)[2], FUN = "*")/j
E13 <- outer(rowSums(M13),colSums(M13)[2], FUN = "*")/j
E23 <- outer(rowSums(M23),colSums(M23)[2], FUN = "*")/j
DB <- sum(abs(M12[,2]-E12),  abs(M13[,2]-E13),  abs(M23[,2]-E23))/j
#cat("  E12 :",  E12 ,"\n")
#cat("  E13 :",  E13 ,"\n")
#cat("  E23 :",  E23 ,"\n")
```

```
#cat("DA =",DA," DB=",DB,"\n")

#select the next treatment
diff <- DA - DB
if(abs(diff) < 1.0e-5)
{probca <- 0.5} else
  probca <- ifelse(diff>0,pfavour,(1-pfavour))

probraca<-(probca*probra)/(probca*probra+(1-probca)*(1-probra))

if(probtemp<=1-Pu)
{duration[isim]<-rectime[j]
 early[isim]<-1
 death[isim]<-d
 stop<-1
 chose[isim]<-2}# trt A is chosen to be superior trt
if(probtemp>=Pu)
{duration[isim]<-rectime[j]
 early[isim]<-1
 death[isim]<-d
 stop<-1
 chose[isim]<-1}# trt B is chosen to be superior trt

newtreat <- rbinom(1,1,probraca)
#  cat("probca",  probca,"\n")
#cat("probra",  probra,"\n")

# cat("probraca",  probraca,"\n")
treat<- c(treat,newtreat)
if (newtreat==0)
{adasta<-rmultinom(1, size = 1, prob = underp1)
 adast<-sum(adasta*c(1:catnum))
 if(adast==1){
   adaplog<-b01+bt1*newtreat+b11*X1[j]+b21*X2[j]+b31*X3[j]
   adaelamda<-1/exp(adaplog)
 } else if (adast==2){
   adaplog<-b02+bt2*newtreat+b12*X1[j]+b22*X2[j]+b32*X3[j]
   adaelamda<-1/exp(adaplog)
 } else if (adast==3){
   adaplog<-b03+bt3*newtreat+b13*X1[j]+b23*X2[j]+b33*X3[j]
   adaelamda<-1/exp(adaplog)
 } else if (adast==4){
   adaplog<-b04+bt4*newtreat+b14*X1[j]+b24*X2[j]+b34*X3[j]
   adaelamda<-1/exp(adaplog)
 }
```

```
     Tnew<-rexp(1,adaelamda)
     mat012 <- newmat012
     mat013 <- newmat013
     mat023 <- newmat023

    }# loop trt A
    if (newtreat==1)
    {adasta<-rmultinom(1, size = 1, prob = underp2)
     adast<-sum(adasta*c(1:catnum))
     if(adast==1){
       adaplog<-b01+bt1*newtreat+b11*X1[j]+b21*X2[j]+b31*X3[j]
       adaelamda<-1/exp(adaplog)
     } else if (adast==2){
       adaplog<-b02+bt2*newtreat+b12*X1[j]+b22*X2[j]+b32*X3[j]
       adaelamda<-1/exp(adaplog)
     } else if (adast==3){
       adaplog<-b03+bt3*newtreat+b13*X1[j]+b23*X2[j]+b33*X3[j]
       adaelamda<-1/exp(adaplog)
     } else if (adast==4){
       adaplog<-b04+bt4*newtreat+b14*X1[j]+b24*X2[j]+b34*X3[j]
       adaelamda<-1/exp(adaplog)
     }

     Tnew<-rexp(1,adaelamda)
     mat112 <- newmat112
     mat113 <- newmat113
     mat123 <- newmat123
    }# loop trt B

    st<-c(st,adast)
    TT<-rbind(TT,c(Tnew, Tnew+rectime[j]))

    j<-j+1
  }#end of iterations through patients

  if(j==ntotal+1 & stop==0)
  {
    TTT<-ifelse(TT[,2]<=maxtime, TT[,1], (maxtime+TT[,1]-TT[,2]))
    #na<-sum(ifelse(treat==0,1,0))
    # nb<-sum(ifelse(treat==1,1,0))

    for(cc in 1:catnum)
    {
      upp1[cc]<-gamma[cc]+sum(ifelse(treat==0 & st==cc ,1,0))
      upp2[cc]<-gamma[cc]+sum(ifelse(treat==1 & st==cc ,1,0))
      t0[cc]<-sum(TTT[treat==0 & st==cc])
      t1[cc]<-sum(TTT[treat==1 & st==cc])
      n0[cc]<-length(TT[(TT[,2]<=maxtime & treat==0 & st==cc),1])
      n1[cc]<-length(TT[(TT[,2]<=maxtime & treat==1 & st==cc),1])
    }

    d_final<-n0[1]+n1[1]
    ga1<-alpha+n0
    ga2<-alpha+n1
    gb1<-beta+t0
    gb2<-beta+t1
    #cat(" upp1",   upp1,"\n")
    #cat(" upp2",   upp2,"\n")

    # cat("ga1",  ga1,"\n")
    # cat("ga2",  ga2,"\n")
    # cat("gb1",  gb1,"\n")
    # cat("gb2",  gb2,"\n")


    probtemp<-comp(ga1,ga2,gb1,gb2,upp1,upp2)
    # cat("probtemp",  probtemp,"\n")
    if(probtemp<=1-Pu)
    {
      fchose[isim]<-2}# trt A is chosen to be superior trt
    if(probtemp>=Pu)
    {
      fchose[isim]<-1}# trt B is chosen to be superior trt
    death[isim]<-  d_final
    duration[isim]<-maxtime
  }
  nlength[isim]<-length(treat)
  #treatpro[isim]<-mean(treat)
  imbalance[isim] <- ifelse(newtreat==1,DB,DA)
  patientB_raca[isim]<-length(treat)*mean(treat)
  patientA_raca[isim]<-length(treat)*(1-mean(treat))
}#loop of simulations
chose2<-mean(ifelse(chose==1,1,0))+mean(ifelse(fchose==1,1,0))
chose1<-mean(ifelse(chose==2,1,0))+mean(ifelse(fchose==2,1,0))

avduration<-mean(duration)
avdeath<-mean(death)
meanpatient<-mean(nlength)
earlytermination <- mean(early)
```

```
sdduration<-sd(duration)
sddeath<-sd(death)
sdpatient<-sd(nlength)

number1<-mean(patientB_raca)
number0<-mean(patientA_raca)

sd1<-sd(patientB_raca)
sd0<-sd(patientA_raca)

power <-chose2

PBA<-(number1/meanpatient)*100

cat("Chance for arm A to be selected as the superior treatment=",
 chose1,"\n")
cat("Chance for arm B to be selected as the superior treatment=",
chose2,"\n")
cat("# of patients in arm A=", number0,"\n")
cat("# of patients in arm B=", number1,"\n")
cat("sd of patients in arm A=", sd0,"\n")
cat("sd of patients in arm B=", sd1,"\n")

cat("# of patients in trial =", meanpatient,"\n")
cat("# of death =", avdeath,"\n")
cat("power of the test =", power,"\n")
cat("The probability of early termination =", earlytermination,"\n")
cat("The average length of trial =", avduration,"\n")

cat("The percentage of patients assigned to the better treatment =",
 PBA,"\n")

cat("# of patients =", meanpatient,"\n")
cat("sd of patients =", sdpatient,"\n")
cat("sd of duration =", sdduration,"\n")
cat("sd of death =", sddeath,"\n")


cat("The average degree of imbalance measured at the end of the trial",
mean(imbalance ), "\n")
cat("The sd degree of imbalance measured at the end of the trial",
sd(imbalance ), "\n")
```

References

Asimow, L. and Maxwell (2010) *Probability and Statistics with Applications: A Problem Solving Text.* Winsted, CT: ACTEX.

Bandyopadhyay, U. and Bhattacharya, R. (2006) Adaptive allocation and failure saving in randomised clinical trials. *Journal of Biopharmaceutical Statistics*, **16**, 817–829.

Barbachano, Y., Coad, D. S. and Robinson, D. R. (2008) Predictability of designs which adjust for imbalances in prognostic factors. *Journal of Statistical Planning and Inference*, **138**, 756–767.

Bather, J. (1981) Randomized allocation of treatments in sequential experiment. *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**, 265–292.

Berry, D. (2004) Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science*, **19**, 175–187.

Bewick, V., Cheek, L. and Ball, J. (2004) Statistics review 12: survival analysis. *Critical Care*, **8**, 389–394.

Biswas, A. and Bhattacharya, R. (2012) Response-adaptive designs for continuous treatment responses in phase III clinical trials: A review. *Statistical Methods in Medical Research.*

Brooks, R. J. (1982) On the loss of information through censoring. *Biometrika*, **69**, 137–144.

Brutti, P., Santis, F. D. and Gubbiotti, S. (2008) Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine*, **27**, 2290–2306.

Chang, M. (2008) *Design Theory and Implementation Using SAS and R.* Boca Raton, FL: Chapman and Hall/CRC Press.

Cheng, Y. and Shen, Y. (2005) Bayesian adaptive designs for clinical trials. *Biometrika*, **92**, 633–646.

Chow, S., Chang, M. and Pong, A. (2005) Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, **15**, 575–591.

Ciolio, J., Zhao, W., Martin, R. and Palesch, Y. (2011) Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. *Clinical Trials*, **32**, 250–259.

Clarke, B. and Yuan, A. (2006) Closed form expressions for Bayesian sample size. *Annals of Statistics*, **34**, 1293–1330.

Cook, T. and DeMets, D. (2008) *Introduction to Statistical Methods for Clinical Trials.* Boca Raton, FL: Chapman & Hall/CRC.

Dragalin, V. (2006) Adaptive designs: Terminology and classification. *Drug Information Journal*, **40**, 425–435.

Dupont, W. and Plummer, W. (1990) Power and sample size calculations: A Review and Computer Program. *Controlled Clinical Trials*, **11**, 116–128.

Efron, B. (1971) Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.

Emerson, C. S., Kyle, D. R. and Emerson, S. S. (2011) Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in Medicine*, **30**, 1199–1217.

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2007) Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*, **26**, 5047–5080.

Faraway, J. (2006) *Extending the Linear Model with R: generalized linear, mixed effects and nonparametric regression models.* Boca Raton, FL: Chapman & Hall/CRC.

Frane, J. W. (1998) A method of biased coin randomization, its implementation, and its validation. *Drug Information Journal*, **32**, 423–432.

Gail, M. (1985) Applicability of sample size calculations based on a comparison of proportions for use with the log-rank test. *Controlled Clinical Trials*, **6**, 112–119.

Gaydos, B., Lewis, R., Maca, J., Pinheiro, J., Pritchett, Y. and Krams, M. (2009) Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal*, **43**, 539–556.

Green, S., Crowley, J. and Benedetti, J. (2008) *Clinical Trials in Oncology*. Boca Raton, FL: Chapman & Hall/CRC.

Greenhouse, J. and Wassermann, L. (1995) Robust Bayesian method for monitoring clinical trials. *Statistics in Medicine*, **14**, 1379 – 1391.

Hagino, A., Hamada, C., Yoshimura, I., Ohashi, Y., Sakamoto, J. and Nakazato, H. (2004) Statistical comparison of random allocation methods in cancer clinical trials. *Controlled Clinical Trials*, **25**, 572–584.

Hogg, R., McKean, J. and Craig, A. (2013) *Introduction to Mathematical Statistics.* 7th edtion. Boston, MA: Pearson.

Hu, F., Hu, Y., Ma, Z. and Rosenberger, W. (2014) Adaptive randomization for balancing over covariates. *Computational Statistics*, **6**, 288–303.

Hu, F. and Rosenberger, W. (2006) *The Theory of Response-Adaptive Randomization in Clinical Trials.* New York: John Wiley & Sons Inc.

Huang, X., Ning, J., Li, Y., Estey, E., Issa, J. and Berry, D. (2009) Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in Medicine*, **28**, 1680–1689.

Inoue, L., Berry, D. A. and Parmigiani, G. (2005) Relationship between Bayesian and frequentist sample size determination. *The American Statistician*, **59**, 79–87.

Jiang, F., Lee, J. J. and Muller, P. (2013) A Bayesian decision-theoretic sequential response-adaptive randomization design. *Statistics in Medicine*, **32**, 1975–1994.

Kalish, L. and Begg, C. (1985) Treatment allocation methods in clinical trials: A review. *Statistics in Medicine*, **4**, 129–144.

Kleinbaum, D. G. and Klein, M. (2005) *Survival analysis: a self-learning text.* 2nd edition. New York: Springer.

Korn, E. and Freidlin, B. (2011) Outcome-adaptive randomisation: Is it useful? *Journal of Clinical Oncology*, **29**, 771–776.

Lakatos, E. (1988) Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, **44**, 229–241.

— (2002) Designing complex group sequential survival trials. *Statistics in Medicine*, **21**, 1969–1989.

Montgomery, D. (2001) *Design and Analysis of Experiments.* 5th edition. New York: John Wiley & Sons Inc.

Mukhopadhyay, N. (2000) *Probability and Statistical Inference.* Boca Raton, FL: Chapman and Hall/CRC Press.

Myers, R. H., Montgomery, D. C., Vining, G. G. and Robinson, T. (2010) *Generalized Linear Models: with Applications in Engineering and the Sciences.* New York: John Wiley & Sons Inc.

Ning, J. (2009) `https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?SoftwareId=82`. Accessed:August 30, 2011.

Ning, J. and Huang, X. (2010) Response-adaptive randomization for clinical trials with adjustment for covariate imbalance. *Statistics in Medicine*, **29**, 1761–1768.

Parnell, I. (2002) *Use of decision analysis to design a habitat restoration experiment.* Master's thesis, Resource and Environmental Management, Simon Fraser University.

Peto, R., Pike, M. C., Armitage, P., Breslow, N., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II analysis and examples. *British Journal of Cancer*, **35**, 1–39.

Pocock, S. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.

— (1993) *Clinical trials:*A *Practical Approach.* New York: John Wiley & Sons Inc.

Pocock, S. and Simon, R. (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, **31**, 103–115.

Rosenberger, W. and Hu, F. (2004) Maximizing power and minimizing treatment failures in clinical trials. *Clinical Trials*, **1**, 141– 147.

Rosenberger, W. and Lachin, J. (1993) The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials*, **14**, 471–484.

— (2002) *Randomization in Clinical Trials: Theory and Practice.* New York: John Wiley & Sons Inc.

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

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

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

Schneeweiss, S. (2006) Sensitivity analysis and external adjustment for unmeasured confounders. *Pharmacoepidemiology and Drug Safety*, **15**, 291–303.

Shih, J. H. (1995) Sample size calculation for complex clinical trials with survival endpoint. *Controlled Clinical Trials*, **16**, 395–407.

Sylvester, R. J. (1988) A Bayesian approach to the design of Phase II clinical trials. *Biometrics*, **44**, 823–836.

Taves, D. (1974) Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology Therapeutics*, **15**, 44–453.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M., Giangregorio, L., Dennis, B., Kosa, D., Debono, V., Dillenburg, R., Fruci, V., Bawor, M., Lee, J., Wells, J. and Goldsmith, C. H. (2013) A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology*, **13**, 92–104.

Thall, P. and Wathen, J. (2005) Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine*, **24**, 1947–1964.

Thompson, W. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

Viel, F., Pobel, D. and Carre, A. (2007) Incidence of leukaemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis. *Statistics in Medicine*, **14**, 2459 – 2472.

Wang, D. and Bakhai, A. (2006) *Clinical Trials.* Illinois: Remedica.

Wunder, C., Kopp-Schneider, A. and Edler, L. (2012) An adaptive group sequential phase II design to compare treatments for survival endpoints in rare patient entities. *Journal of Biopharmaceutical Statistics*, **22**, 294–311.

Yin, G. (2012) *Clinical Trial Design.* New York: John Wiley & Sons Inc.