RESPONSE ADAPTIVE CLINICAL TRIALS WITH CENSORED LIFETIMES

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Xun Dong

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
Room 142 McLean Hall
106 Wiggins Road
Saskatoon, Saskatchewan
CANADA
S7N 5E6

Abstract

We have constructed a response adaptive clinical trial to treat patients sequentially in order to maximize the total survival time of all patients. Typically the response adaptive design is based on the urn models or on sequential estimation procedures, but we used a bandit process in this dissertation. The objective of a bandit process is to optimize a measure of sequential selections from several treatments. Each treatment consist of a sequence of conditionally independent and identically distributed random variables, and some of these treatment have unknown distribution functions. For the purpose of this clinical trial, we are focusing on the bandit process with delayed response. These responses are lifetime variables which may be censored upon their observations. Following the Bayesian approach and dynamic programming technique, we formulated a controlled stochastic dynamic model. In addition, we used an example to illustrate the possible application of the main results as well as R to implement a model simulation.

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor Professor M. G. Bickis. This dissertation would not have been written without his advice and assistance. In addition, I want to thank Xikui Wang, for the inspiration and direction his dissertation gave me.

I would like to thank the Department of Mathematics and Statistics for Providing me scholarship.

Finally, I would like to thank my family for the encouragement and support they have given me as I complete my dissertation and degree.

Dedicated to my parents:

Xun Qing-Feng

Xia Yu-Feng

Tian Xiao-Yi

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

It is widely accepted by the scientific community that a clinical trial is the most reliable and efficient way to evaluate the efficacy of new medical interventions, therapeutic or prophylactic. Randomization has always been an essential feature for designing such clinical trials. It prevents selection bias and insures against accidental bias. The traditional randomization has been regarded as the gold standard for clinical trials. In particular, the permuted block design is the most popular randomization procedure applied in practice. However, as an experiment on human subjects, maintaining the balance between collective ethics and individual ethics is difficult in clinical trials. In some desperate clinical trials, such as cancer, the patient is suffering from a serious, acute, and potentially debilitating or terminal illness. The collective ethics fall into a disadvantaged position, and the traditional balanced randomization becomes unavailable because of the unjustifiable sacrifice of individual ethics. Instead, response adaptive randomization is ethically justified and morally required[23].

An important class of clinical trial designs is adaptive randomization, which has been a fruitful area of research. Adaptive randomization refers to any scheme in which the allocation probability changes according to response of treatment in the trial. In the past decade many books and top statistical journals have studied this subject. Two recent book chapters detail the modern statistical theory and practice of adaptive randomization: Pong and Chow[26] and Hinkelmann[18]. Many recent developments also focused on providing a template for the selection of appropriate procedures and guidance for their use in practice. For example, the U.S. Food and Drug Administration (FDA, 2010) drafted a guide to regulatory aspects of adaptive randomization designs. A response adaptive randomization can be considered as a data dependent design of clinical trials. The goal is to improve the clinical trials' efficiency and ethics without undermining the validity and integrity of the clinical research[43]. Typically the response adaptive design is based on urn models or on sequential estimation procedures[15], but a more promising approach is to use the decision theoretic model of bandit process[4]. The advantage of the bandit model is that it is deterministic and achieves an optimal solution[7]. For more details on adaptive designs applied in clinical trials, Yi and Wang[45] did a general review of methods, issues and inference procedures of response. In addition, a prototypical example will be discussed later.

While response adaptive randomization procedures are not appropriate in clinical trials with a limited recruitment period and outcomes that occur after all patients have been randomized, there is no reason why response adaptive randomization cannot be used in clinical trials with moderate delayed responses[32]. The delayed response just represents the delayed times for the responses of the treatments in clinical trial. In addition, there is much literature showing that, under widely applicable conditions, large sample are unaffected by delayed responses for the: doubly-adaptive biased coin design[17], urn models[5, 14, 48], and bandit model[40]. From a practical perspective, there is no difficulty in applying delayed responses to the response adaptive randomization procedure provided that responses become available during the recruitment and randomization period. The information update can also be made after groups of patients or individual patients have responded. In this dissertation, we will construct a controlled stochastic dynamic model to discover some properties of the optimal deterministic strategy for the bandit process. Our model comes from Wang's[40] bandit model modified. After deriving our main results, we will also highlight a possible application for designing response adaptive clinical trials. Based on this application, we redid Wang's simulations and found there are some initialization problems in his program. In Wang's PhD thesis[40], he simulated *two patients case, three patients case*, and *fourteen patients case*, all these simulations were programmed in Fortran[12]. In my simulation, I redesigned the program by R[33] and included many more different situations, such as different expect arrival time and different known expect survival time. More details are fully investigated in our follow up research.

1.2 Prototypical example

Consider a clinical trial in which two different treatments A and B are available for a disease (such as cancer, for example). Patients with this disease are recruited to participate in the trial. The patients' actual lifetimes after treatment is our primary concern, and our objective is to maximize the measure of the total of the patients' expected lifetimes after treatment. The assumption is that each patient is treated only with one treatment. If supposedly treatment A is a new medical intervention, then its effectiveness is unknown. After treatment A, all patients' lifetimes will follow the same lifetime distribution, which is unknown and is to be estimated. On the other hand, treatment B is standard. That is to say, it has been investigated before and has well known effectiveness. The patients' expected lifetime is known after treatment B. After the treatment, the patients are closely followed by the investigator, and their responses are available at any time. That is to say, the patients' actual lifetimes after treatment have become the focal point. After observing the condition of the patients, the efficacy of the treatment can be inferred.

Assume that there are two independent treatments A and B. Since they are independent, the observations made during one treatment will not provide any information on the other. The patients are treated

sequentially one by one, and the time frame between patients is random. We can assume that the time frame between the treatments of two consecutive patients follow the same arrival distribution and are independent of each other. The effectiveness of treatment A is unknown, but we assume that prior information on treatment A is available. A clinical trial is a responsible and ethical experiment on a human being, so the patients' interests are more important than the doctors'. Hence we should always analyze the observations from previously treated patients on treatment A before a new patient will be treated.

Suppose that at the time a patient is to be treated, the expected lifetime is $E(X \mid A)$ with treatment A, and $E(X \mid B)$ with treatment B, where X is the residual survival time. If treatment A appears to be no worse than treatment B in terms of expected lifetime of the patients (i.e.$E(X \mid A) \geq E(X \mid B)$), then we shall treat this patient with treatment A. On the other hand, if treatment B appears to be better than treatment A in terms of expected lifetime (i.e.$E(X \mid A) \leq E(X \mid B)$), then we shall allocate treatment B. In order to receive a higher expected lifetime for the present patient, we should allocate treatment B. However, the expectation may change as information is updated. If we would allocate treatment A to the patient, although we receive a shorter expected lifetime for the present patient, the observation from this patient will provide information on the efficacy of treatment A in the future. This will improve our understanding of treatment A, and may lead to higher expected lifetimes for future patients.

The best allocation of treatment to a patient is the biggest problem we encountered in this investigation. The allocation depends on the nature and amount of information available, the number of patients to be treated in the future, and the objectives of the trial. The attitude of what constitutes relevant information, the approach of utilizing the information, and the type of design we choose will also influence our allocation of the treatment. Since the patient's lifetime is our primary concern, another difficulty comes from the nature of the available information. In clinical trials, the term *censoring* is used to refer to mathematically removing a patient from the survival curve at the end of their follow-up time due to incomplete information because we do not know the actual lifetime but only a lower bound[25]. After the treatment, a patient's lifetime is continuous, and its observation may be censored at any time. If at a particular time, all patients previously treated with treatment A are dead, then complete information is available about the effectiveness of treatment A. Otherwise, only censored observations of some of these patients' lifetimes are available. Then we would be facing the problem of missing information about the effectiveness of treatment A. In this case, allocation of treatment is a difficult decision. It is practically impossible to wait for complete observations of all previously treated patients' lifetimes before treating the next patient. The doctors cannot wait, neither can the patients, because delayed treatment could cause death, or worsen the sickness. Therefore, the allocation has to be done before complete information is available, and this involves the statistical analysis of censored lifetimes.

The goal of this dissertation is to formulate and analyze sequential treatments with censored lifetimes. We

will review relevant concepts and issues in the subjects of clinical trials, sequential selections of treatments and statistical analysis of lifetimes.

1.3 Clinical trials

According to Lee[21], *Clinical trials* are prospective studies that, under some pre-specified conditions, evaluate the effect of medical interventions in humans. Since the 19th century, clinical trials have become a standard and an integral part of modern medicine. Clinical trials can be used as a tool to better understand human disease. A properly planned and executed clinical trial is the most definitive tool for evaluating the effect and applicability of new treatment modalities. Clinical trials have been studied extensively in the statistical literature; see Piantadosi[24] for methodological perspectives of clinical trials; see DeMets[9] for current developments in clinical trials; and further investigated in D'Agostino[10], a panel discussion on the future of clinical trials. In order to better understand clinical trials, we will introduce basic concepts below.

The *patient* denotes an individual enrolled in clinical trials. Patients usually are divided into separate groups to receive different treatments. Some will be allocated to the treatment group, while others will be allocated to a comparison group.

The *treatment* in a clinical trial can be considered as any medical experiment on the patient. The experiment may be administering a drug, performing an operation, or simply giving the patients a placebo. The *test treatment* is a medical intervention to be evaluated in the trial for its efficacy. A clinical trial may involve one or more test treatments. The *control treatment* is a standard treatment used for comparison with the test treatments.

The basic *goal* of a clinical trial is to estimate the efficacy of treatment and to treat the patient as effectively as possible. In order to assess the efficacy from different treatments, the designer normally allocates the matched stratified design to all groups.

A clinical trial normally involves four phases: *design, execution, analysis*, and *reporting*. Our research concentrated on phase one, which are designed to establish the basic safety and efficacy of a test treatment. A successful clinical trial needs to meet basic prerequisites, and the objectives must be precise and clear. If a design has several purposes, the most important one should be explicitly defined.

Since the patients' responses are biological, they will inevitably vary if the same treatment is applied to the different patients with the same disease, or to the same patient repeatedly. Therefore, the variation in responses are due to the treatments and individual uncertainties. In order to avoid this as much as possible,

the criteria for recruiting patients into clinical trials should also be well defined.

1.4 Adaptive randomization

Randomization is a fundamental component of a well conducted clinical trial. Randomization ensures the comparability of treatment groups and mitigates the selection bias in the design. It plays a critical role among advances in the history of medical research. The traditional randomization has been regarded as the gold standard for clinical trials, such as permuted block randomization in which the number of subjects in each block are specified, and subjects are allocated randomly within each block. It is the most popular randomization procedure applied in practice. However, it may not be optimal in many clinical trials where multiple experimental objectives are pursued. Another important class of randomization applied in clinical trials is *adaptive randomization*. It is a change in allocation probabilities during the course of the trial to promote multiple experimental objectives, while protecting the study from bias and preserving inferential validity of the results[32]. Most clinical trials are multi-objective, and it is hard to find a single design criterion that describes all objectives. However, adaptive randomization can be considered as a solution to an optimization problem which accommodates several selected design criterion. In the design phase, some of the objectives will depend on model parameters which are unknown. During the course of the trial, adaptive randomization is useful and sometimes imperative to redesign the trial to achieve these objectives. In addition, the traditional randomization is not applicable in some desperate clinical trials since there is an unjustifiable sacrifice of individual ethics. Therefore, adaptive randomization has an inherent advantage and our model based on the *response-adaptive randomization*. It is a randomization procedure that uses past treatment assignments and patient responses to select the future treatment assignments, with the objective to maximize power and minimize the probability of treatment failures.

Response-adaptive randomization has been studied extensively in statistical literature. Pioneering works can be traced to Thompson[38] and Robbins[28]. Extensive research has been done by Anscombe[1] and Colton[6], with more recent developments in this area by Rosenberger and Lachin[30]. A more detailed history of the subject is discussed in Rosenberger, Sverdlov and Hu[32]. The most famous nonrandomized response-adaptive methodology is *play-the-winner rule*[46], in which a success with a treatment leads to the assignment of the next patient to the same treatment, and failure would assign the next patient to the opposite treatment. The first randomized response-adaptive methodology was designed by incorporating randomization into the play-the-winner rule, and was named the *randomized play-the-winner* rule[39]. With this approach, the patient's treatment assignment keeps the spirit of the play-the-winner rule, but the patient can be modelled by drawing a ball from an urn, and the urn composition is updated based on the previous responses so that there are more balls corresponding to the better treatment. Based on the randomized play-the-winner rule, most research on response-adaptive randomization is based on the stochastic processes

such as *urn models*[2] and *bandit processes*[3]. Modern research on response-adaptive randomization has focused on the development of optimal response-adaptive randomization procedures that maintain or increase power over traditional balanced randomization designs, and minimize expected treatment failures for clinical trials with two or more treatment groups[32]. In the past decade, the most representative advance on response-adaptive randomization is *drop-the-loser rule*[19] which is a fully randomized procedure based on the urn model with minimal variability. Generalizations of the drop-the-loser rule can be found in Sun et al.[35] and Zhang et al.[48]. Although the urn model received the most attention in the latter part of the 20th century[31], we will focus on the bandit model in this dissertation, and will discuss this in more detail in the next section.

1.5 Bandit processes

In clinical trials, a *sequential design* is a method for minimizing the number of patients on the inferior treatment, and analyzing and monitoring data continually as it becomes available[26]. The sequential design normally takes advantage of accruing information to optimize experimental objectives, and assumes that the only decision to be made is whether the trial should continue or be terminated depending on the effect. *Bandit processes* are sequential decision problems with successive selections from different arms, which represent treatments in clinical trials. Each arm of the bandit process represents a treatment which consists of a sequence of random responses given by a conditionally independent and identical distribution with known or unknown parameters. At each stage, the designer must decide which arm to apply next. Selecting an arm means choosing a treatment for observation. This means assigning a treatment to the next patient in the clinical trial. The choice normally depends on a trade-off between *exploiting arms* with utility that appear to be doing well, and *exploring arms* that might potentially be optimal in the future, but appear to be weak at present. The goal of the bandit process is to maximize the objective function of responses from all selections. The bandit process was first posed by Thompson[38] but received no attention until it was studied by Robbins[28]. Following the research of Berry and Fristedt[3] and Gittins[13], the bandit process has had an extensive development. Gittins went on to redefine the bandit process as the semi-Markov decision process. Typically, the bandit problem is applied in a *multi-armed bandit process*, which is a sequential experiment with the goal of achieving the largest possible expect value from a distribution with unknown parameters[36]. The multi-armed bandit process has at least two arms. For example, a simple case of a *one-armed bandit process* can be a sequential experiment with two arms, one with a known distribution and the other with an unknown distribution. According to our prototypical example, the one-armed bandit process is very suitable for our design ideas. We will do a detailed discussion of its model, states and strategies in Chapter 3.

1.6 Statistical analysis of lifetimes

Mathematically, a *lifetime* is a non-negative real-valued random variable in a probability space. The statistical analysis of lifetimes is called *survival analysis*, which is defined broadly as the time to the occurrence of a given event. This event can be the development of a disease, response to a treatment, relapse, or death[20]. In clinical trials, we normally want to compare the lifetimes in two groups to see whether the lifetimes of individuals in one group are systematically longer than those in the other. However, lifetimes of individuals in a single group can also be examined to discover other influences. There are many methodologies that analyze lifetime data, such as various parametric models, nonparametric models, associated distribution-free methods, graphic procedures, and life tables. The basic design idea is how to specify models in different circumstances to represent lifetime distributions and to make statistical inference on the basis of these models. The lifetime distribution may be described in various ways, such as the distribution function, the survival function, the hazard function, or the mean residual lifetime function. In this dissertation, we mainly focus on the survival lifetime function.

For a random lifetime variable T, the *distribution function* is

$$F(t) = P(T \leq t),$$

the *survival function* is

$$S(t) = P(T > t) = 1 - F(t),$$

and the *mean residual lifetime function* is

$$M(t) = E(T - t | T \geq t).$$

If the distribution is absolutely continuous with density function $f(t) = F'(t)$, the *hazard function* $h(t)$ is defined as the instantaneous rate of death at time t, given that $T \geq t$, that is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

For a discrete random lifetime variable T, which takes on values $0 \leq t_1 < t_2 < ...$, let the *probability mass function P(t)* be given by

$$P(t_j) = P(T = t_j), j = 1, 2, ...$$

In this case the hazard function $h(t)$ is defined by

$$h(t_j) = P(T = t_j | T \geq t_j) = \frac{P(t_j)}{S(t_j)}, j = 1, 2, ...$$

These functions all give mathematically equivalent specifications of the distribution of the lifetime variable T. The hazard function is often the most useful representation because it describes how the instantaneous probability of death for an individual changes with time. Especially the case when there are special reasons

to restrict consideration to models with non-increasing hazard functions, or with hazard functions having some specific characteristics. Qualitatively, hazard functions are sometimes classified as monotone increasing functions, monotone decreasing functions, bathtub-shaped, or other shaped. There are other distributions that occupy a central role in the parametric models used to analyze lifetime data, such as exponential distributions, Weibull distributions, gamma distributions, and log-normal distributions. For more details see Pappas, Adamidis and Loukas[27].

In the real analysis of lifetime data it is hard to observe the full lengths of the individuals' lifetimes. In most clinical trials, not all patients are dead before the end of the trial. If we can only get partial information on the lifetimes of some individuals, then such datas are censored. The main problem in censored lifetimes is to determine the sampling distribution and corresponding likelihood function for the process, and to determine the properties of statistical methods derived from this.

Censoring can be classified in several ways[40]. An observation of the lifetime is said to be *right censored at time* O if the exact value of the lifetime is not known at time O but is known to be greater than or equal to O. On the other hand, an observation of the lifetime is said to be *left censored at time* O if the exact value of the lifetime is not known at time O but is known to be less than or equal to O. Right censoring is common in lifetime data analysis. If patients are still alive at the time of observation then their lifetimes are right censored. Left censoring usually occurs if observation does not start immediately and some individuals have already died before the observation starts.

Suppose there are n individuals, such that the ith individual's lifetime T_i is observed only if $T_i \leq O_i, i = 1, ..., n$. Therefore, O_i may be the censoring time for individual i, and we call it a *potential censoring time* if it is considered without regards to whether censoring or death occurs. Let us consider the continuous lifetime distribution case and suppose that the censoring times $O_1, .., O_n$ are fixed. We also assume that the T_i are independent and identically distributed with the density function $f(t)$ and survival function $S(t)$. Then, for individual i, we can only observe

$$t_i = \min(T_i, O_i),$$

and the indicator function

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq O_i, \\ 0 & \text{if } T_i > O_i, \end{cases}$$

where $\delta_i = 1$ indicates that death is observed, and $\delta_i = 0$ indicates that censoring occurs.

Note that t_i is a mixed random variable with a continuous and a discrete component, and δ_i is a discrete random variable. The joint density function of t_i and δ_i can be expressed as

$$f(t_i)^{\delta_i} S(O_i)^{1-\delta_i}.$$

If the pairs (t_i, δ_i) are independent, the likelihood function becomes

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(O_i)^{1-\delta_i}.$$

Based on our prototypical example, we assume the lifetime distribution is an exponential distribution. In this case,

$$S(t) = e^{-\theta t}$$

and

$$L = \theta^r e^{-\theta(\sum_{i=1}^{n} t_i)}$$

where $r = \sum_{i=1}^{n} \delta_i$ is the total observed number of deaths, and $\sum_{i=1}^{n} t_i$ is the total observed lifetime.

1.7 A brief summary

Through the above general introduction, we have constructed a prototypical example, and specified assumptions for this example. We also reviewed relevant background information in the areas of clinical trials, randomization, sequential method and lifetime data analysis. These build the foundation of our dissertation and the more detailed discussion in the following chapters. In Chapter 2, we will discuss the basic ingredients of general Bayesian approach for clinical trials. This includes fundamental elements, Bayesian approach, and adaptivity applied in clinical trials. All of these are useful for later studies. In Chapter 3, we will discuss the sequential selection of treatment with delayed response, and will give a full detailed formula on a special controlled stochastic processes. We will derive the optimality equations of the bandit process and will list the theoretical properties. In Chapter 4, we will focus on detailed results and from that derive the optimality equations in a prototypical example. Finally, we will present our conclusions and discuss prospects for future research.

Chapter 2

Bayesian approach for clinical trials

In this chapter, we progressively provide an overview of Bayesian modelling and inference, to explain why the Bayesian approach is important in clinical trials, and how it works to support our special controlled stochastic processes. Typically, the *frequentist* is the standard statistical approach to designing and analyzing in clinical trials. However depending on the primary purpose, we focus on an alternative approach called the *Bayesian approach*, which began with a posthumous publication in 1763 by Thomas Bayes, an English clergyman who lived from 1702 to 1761.

2.1 Bayesian approach

We define a *Bayesian approach* as the explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation and reporting of a clinical trial[34]. Such a perspective can be more *flexible* than traditional methods in that it can adapt to each unique situation, more *efficient* in using all available evidence, more *useful* in providing predictions and inputs for making decisions for specific patients, for planning research or for public policy, and more *ethical* in both clarifying the basis for randomization and fully exploiting the experience provided by past patients.

Generally, the Bayesian approach has two major advantages[8]. Firstly, experiments can be altered in midcourse, disparate sources of information can be combined, and expert opinion can play a role in inferences. Secondly, the Bayesian approach can be decision-oriented, with experimental designs tailored to maximize objective functions, such as overall public health benefit. So the basic idea of Bayesian approach in clinical trials is reasonably straightforward. In the next section, we will discuss the fundamental elements used in Bayesian approach.

2.2 Fundamental elements

In *frequentist* statistics the observation y is random, with a density or probability mass function given by $f(y;\theta)$, and the parameter θ is treated as a fixed unknown value. On the other hand, in Bayesian statis-

tics y and θ are both regarded as random variables, with joint density given by $\pi(\theta)f(y;\theta)$, where π is the prior density on θ. By modelling both the observed data and any unobserved data as random variables, the Bayesian approach to statistical analysis provides a complex model with external knowledge, expert opinion, or both. Before we introduce the technical details of the Bayesian approach we will first discuss some details of likelihood, prior distribution and posterior distribution, which play a central role in the arguments of this dissertation. The key conceptual point is the way that the *prior* distribution on the unknown parameter θ is updated, on observing the realized value of the data y, to the *posterior distribution*, via Bayes' law.

2.2.1 The likelihood function

The likelihood function of a set of parameters θ given some observed values y is equal to the probability density function of those observed values given those parameter values, the function associates the values $f(y;\theta)$ to each θ. The likelihood function is denoted by $L(\theta;y)$. We mostly work with its logarithm $l(\theta;y)$, often abbreviated to $l(\theta)$. The log form is due to convenience, in particular because the likelihood function will often be a product of component terms. Occasionally we work directly with the likelihood function itself. By using a particular value θ, the larger the value of L the greater are the chances associated to the observation under consideration. Therefore, by fixing the observed values y and varying parameter θ, we can observe the likelihood value of each θ. This is sometimes elevated into a principle called *the likelihood principle*.

The likelihood function leads to the likelihood principle, which states that by making inferences or decisions about θ after y is observed, all relevant experimental information is contained in the likelihood function for the observed y. This principle draws a clear line that separates the inference schools. On one side, the Bayesian and likelihood approaches do not violate this principle, and on the other side the frequentist approach is based on the probabilities implied by the sampling distribution of y.

2.2.2 The prior distribution

In a Bayesian analysis, prior information about a parameter θ is assessed as a probability distribution on θ. This distribution depends on the assessor and can be calculated any time when the assessor has a precise enough opinion. Since posterior distributions depend on prior distributions, we can use prior distributions to drive Bayesian analysis at any time when the assessor has an opinion and therefore it influences conclusions. When a person's opinion about a parameter includes a broad range of parameter possibilities, a diffuse prior distribution may be chosen. In some cases a prior distribution is so diffuse that it does not integrate to 1, which we call an *improper prior*. In other cases, opinions may allow for only a few values of the parameter. And the most extreme case would be a prior that places all its mass on one value.

A typical prior distribution for which it is simple to assess influence is named *conjugate prior*. When a conjugate prior distribution is available, conjugate prior property can be used to obtain a posterior density of the same form as prior density but with different parameters. When this happens, the common parametric form of the prior and posterior is called a *conjugate prior family*. Conjugate families are very convenient and allow a variety of shapes wide enough to capture our prior beliefs. There is no universal law that says we must use a conjugate prior. In non-conjugate cases the posterior distribution must be computed numerically. Therefore, in cases where we can find a conjugate family, it is very common to use it. We will give a detailed example combined with Bayes' theroem in the next section.

2.2.3 Bayes' Theorem

Bayes' theorem is the basic tool of Bayesian analysis. To specifying a likelihood distributional model $f(y \mid \theta)$ for the observed data $y = (y_1, ..., y_n)$ given a vector of unknown parameters $\theta = (\theta_1, ..., \theta_k)$, suppose that θ is a random quantity sampled from a *prior* distribution $\pi(\theta \mid \lambda)$, where λ is a vector of hyper-parameters. The prior distribution tells us how to update beliefs based upon observations. To update beliefs about an unknown parameter θ, Bayes' theorem is used to calculate the posterior probability of the unknown parameter. If λ is known, inference concerning θ is based on its *posterior distribution*,

$$p(\theta \mid y, \lambda) = \frac{p(y, \theta \mid \lambda)}{p(y \mid \lambda)} = \frac{p(y, \theta \mid \lambda)}{\int p(y, \theta \mid \lambda)\, d\theta} = \frac{f(y \mid \theta)\pi(\theta \mid \lambda)}{\int f(y \mid \theta)\pi(\theta \mid \lambda)\, d\theta}. \tag{2.1}$$

Since, in practice, λ will not be known, a *hyperprior* distribution $h(\lambda)$ will be required in this approach, and (2.1) will be replaced with

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)} = \frac{\int f(y \mid \theta)\pi(\theta \mid \lambda)h(\lambda)\, d\lambda}{\iint f(y \mid \theta)\pi(\theta \mid \lambda)h(\lambda) d\theta d\lambda}. \tag{2.2}$$

This multi-stage approach is called *hierarchical modelling*, a subject to which I will give more details in the next section. Rewriting Bayes' theorem: "posterior probability $\propto$ likelihood $\times$ prior probability",
where $\propto$ means proportional to. Again, Bayes' theorem provides a formalism for learning clinical trials: the prior represents the previous patients' historical information what was thought before seeing the next patient's data, the likelihood represents the patients' data now available, and the posterior represents what is

thought given both prior information and the data just observed. Continual and instantaneous updating can occur as today's posterior becomes tomorrow's prior. To better understand this process, we use an example to explain it.

Example 2.1:

Suppose we have observed a normal distribution $Y \sim N(\theta, \sigma^2)$ with σ^2 known, so that the likelihood $L(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(y-\theta)^2}{2\sigma^2}), y \in R, \theta \in R$, and $\theta^2 > 0$. If we specify the prior distribution as $\pi(\theta) = N(\theta|u, \tau^2)$, then from (2.1) we can get the posterior as

$$\begin{aligned} p(\theta|y) &= \frac{N(\theta|u,\tau^2)N(y|\theta,\sigma^2)}{p(y)} \\ &= N\left(\frac{\sigma^2}{\sigma^2+\tau^2}u + \frac{\tau^2}{\sigma^2+\tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right). \end{aligned}$$

Consider the more realistic case where the sample variance σ^2 is unknown. And let $h = 1/\sigma^2$ have a $Gamma(a, \beta)$ prior with probability density function

$$p(h) = \frac{\beta^a}{\Gamma(a)} h^{a-1} e^{-h\beta}, h > 0.$$

Since the likelihood for any one observation y_i is still

$$f(y_i|\theta, h) = \frac{h^{1/2}}{\sqrt{2\pi}} e^{-\frac{h}{2}(y_i-\theta)^2},$$

the posterior of h is proportional to the product of the full likelihood and the prior,

$$\begin{aligned} p(h|y,\theta) &\propto \left[\prod_{i=1}^{n} f(y_i|\theta,h)\right] \times p(h) \\ &\propto h^{\frac{n}{2}} e^{-h/2 \sum_{i=1}^{n}(y_i-\theta)^2} \times h^{a-1} e^{-h\beta} \\ &\propto h^{\frac{n}{2}+a-1} e^{-h[\beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2]}, \end{aligned}$$

where in all three steps we have absorbed any multiplicative terms that do not involve h into the unknown normalizing constant. Looking again at the form of the gamma prior in

$$p(h) = \frac{\beta^a}{\Gamma(a)} h^{a-1} e^{-h\beta}, h > 0.$$

We recognize this form as proportional to another gamma distribution, namely a

$$Gamma\left(n/2+a, \beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2\right),$$

where the θ is known.

Thus, the posterior for h is available via conjugacy, and it turns out the gamma distribution offers a conjugate prior.

After a brief overview of fundamental elements, we can make a clear idea to get the posterior distribution, and build the Bayesian modelling. More detail about Bayesian model will be presented in the next chapter.

2.3 Adaptivity in clinical trials

A fully Bayesian approach uses the likelihood function, the prior distribution, and a utility structure to arrive at a decision. *The prior distribution* is used to summarize all available information of model parameters before the data is observed. It is combined with the likelihood function using Bayes' Theorem to obtain the posterior distribution. *A loss function* assigns numerical values to the various gains and losses that are obtained from unknown parameters. It essentially determines how to weigh outcomes and procedures. Bayesian statistical decision theory suggests choosing procedures that have high utility when averaged with respect to the posterior.

The fully Bayesian approach is insufficient in everyday practice[8]. Firstly, in a regular trials, there are often multiple decision makers, all of whom have their own prior opinions. Secondly, since the data arrive sequentially over time, at each monitoring point decision makers should use backward induction(a process of reasoning backwards in time, from the end of a situation to determine a sequence of optimal actions.) to check whether to stop or continue the trial, account the information for the next observation, and the cost of obtaining them. Thirdly, the eliciting costs and benefits are hard to calculate in the process. Moreover, it seems somewhat arbitrary that the appropriate scales of losses are hard to work with and lead to decision rules. For the above reasons, fully Bayesian approaches have largely failed to gain a foothold in regulatory work.

In a clinical trial, when matters turn to experimental design, the Bayesian approach is somewhat less controversial. This is because in order to carry out a sample size calculation, the designer must have some prior information regarding the likely effect of the treatment and its variability. More formally, all evaluations at the design stage are integrating over uncertainty in both samples and parameters. In the terminology of Rubin(1984), this double integration is a "Bayesianly justifiable frequentist calculation[29]". More informa-

tion about this can be found in "Bayesian Adaptive Methods for Clinical Trials"[37].

Our purpose in this dissertation is to construct a bandit process using the full Bayesian approach to test censored lifetime in clinical trials. In this chapter we have introduced the basic Bayesian approach to support our model. In the following chapter we will start to discuss a special controlled stochastic process named bandit process with delayed response.

Chapter 3

Bandit process with delayed response

3.1 Introduction

Motivated by our purpose and the prototypical example, we will use a bandit process with delayed response to construct our model in this chapter. A *multi-armed bandit process* is a sequential experiment with the goal of achieving the largest possible reward from a payoff distribution with unknown parameters[36]. Typically, a k-armed bandit process considers sequential selections from $k \geq 1$ stochastic processes. In the past decade, bandit process have been briefly mentioned as a possible design of clinical trials. In clinical trials, *arms* usually represents the treatments, and selecting an arm means assigning a treatment to the patient. Selections among the arms are specified by *strategy*, which is a criterion for selection of treatments. A strategy is called *optimal* if it maximizes the expected payoff. In a bandit process, the core content is getting the largest possible reward. The objective from a selection is usually defined as an expected value of the response variable, and considered as the expected value of the patients' residual lifetime in a clinical trial. Based on reward, the *worth of a strategy* is an expected value of all rewards averaging over all possible selections resulting from the strategy. In addition, for the purpose of guaranteeing the finiteness of the expected value from a possibly infinite horizon model, i.e, more patients more wait for present patient. We define a discrete discount sequence $D = (a_1, a_2, ...)$ on future patients, such that $a_n \geq 0$ and $\sum_{n=1}^{\infty} a_n < \infty$. Denote $D_n = (a_1, a_2, ..., a_n, 0, 0, ...), D^m = (a_m, a_{m+1}, ...)$ and $D_n^m = (a_m, ..., a_n, 0, 0, ...), m \leq n$. If $a_n = \beta^{n-1}$ for all n and some $\beta \in (0, 1)$, we say that $D = (a^0, a^1, a^2, ...)$ is a *geometric sequence* and we will set the uniform sequence $D_n = (a_1 = 1, a_2 = 1, ..., a_n = 1, 0, 0, ...)$ in our prototypical example to make the calculation quick and easy. More comprehensive discussions on determining the sequence of factors can be found in Berry and Fristedt[3].

In our model, we will follow the Bayesian approach to employ a Markov decision process. The *Markov decision process* is a framework for modelling decision making in situations where outcomes are partly random and partly under the control of a decision maker. This process is necessary and sufficient for selecting the arms in each strategy. Because the patient feedback is not immediate, all updated information in our model have a delayed response and a lag time is defined for this possibility of delayed responses. If the bandit process has a delayed response, the underlying stochastic process is no longer a Markov process because the future behaviour depends not only on the posterior distribution but also on delayed responses. But if

the responses follow loss-of-memory distributions (like geometric distribution or exponential distribution), the formulated controlled stochastic process becomes a Markov decision process[44]. Corresponding to our prototypical example, we assume that at the beginning time $T_1 = 0$, there is a prior distribution G_1, which is a gamma distribution $g(\eta_1, \tau_1)$ for the hazard rate θ of the unknown treatment. Based on G_1, we start to observe the first patient's lifetime X_{i_1}, where the arm i_1 is selected. After a lag time $y_1 = Y_1(\omega)$, the second patient enters the experiment, then we get the possibly censored observation (Z_{i_1}, I_{i_1}), where I is indicate function. It is applied to update G_1 to a posterior distribution G_2 at the time y_1 using the Bayes' theorem. Then it follows the conjugate family distribution properties, the distribution G_2 is still a gamma distribution, with the parameters (η_2, τ_2) and provides new knowledge about the randomness of the arms... Based on G_2, we select a new arm i_2 arrival at second patient which called stage 2, and start to observe X_{i_2}. The posterior G_2 is now the prior for stage 2 and we continue to observe X_{i_1} if the indicator function $I_{i_1} = 0$. After another delay time y_2 at stage 3, we update the posterior distribution G_3, and select an arm again. Finally, this process of selections is repeated at a sequence of selection time points, and the controlled stochastic processes becomes a Markov decision process.

This chapter is organized as follows: we first introduce some basic definitions of bandit process, then we formulate the model and indicate the objectives, and finally we derive the optimality equations and optimality strategies. Our objective of this chapter is to construct our model and to derive an optimal strategy.

3.2 Arms, censoring times and censored observations

In our model, the *arm* can be consider as a sequence of random variables on a probability space (Ω, F). The arm i consists of a sequence of random response variables $\{X_{(i,n)}; n = 1, 2, ...\}$, which represents n patients' residual lifetime after treatment i. The *response variable* $X_{(i,n)}$ is survival time after some treatment i, and it can be observed at any time after the selection. Since all treatments are allocated at the patients' random arrival times, the arm selection should be made at random selection times. Hence, we define a *sequence of selection times* $\{T_n; n = 1, 2, ...\}$ as $T_1 = 0$, $T_2 = Y_1 + T_1$, $T_3 = Y_2 + T_2$,..., $T_n = Y_{n-1} + T_{n-1}$, where T_i represent arrival times of patients and $\{Y_n; n = 1, 2, ...\}$ is the *sequence of inter arrival times or lag times*. Suppose that only one arm can be selected for observation at each stage. If arm i is selected at stage n, and there is a lag time Y_n before the next stage occurs, then *the possibly censored observation* of response variable at the next stage $n+1$ is $\{Z^i_{n,n+1}, I_{\{X_{(i,n)} \le Y_n\}}\}$, which consists of *censored lifetime* $Z^i_{n,n+1} = \min(X_{(i,n)}, Y_n)$ and its *indicator function*

$$I_{\{X_{(i,n)} \le Y_n\}} = \begin{cases} 1 & \text{if } X_{(i,n)} \le Y_n \\ 0 & \text{otherwise.} \end{cases}$$

If the indicator function $I_{\{X_{(i,n)} \le Y_n\}} = 0$, i.e., during that lag time Y_n, we have a censored observation of $X_{(i,n)}$ at stage $n+1$ and we only can get incomplete information on arm i from $X_{(i,n)}$, and the nth patient is

still living. Otherwise, if $I_{\{X_{(i,n)} \leq Y_n\}} = 1$ indicates that complete observation of $X_{(i,n)}$ is available at stage $n+1$, and the nth patient was dead. Extending to the general case, for any $m \geq n$, the possibly censored observation $\{Z^i_{n,m}, I_{\{X_{(i,n)} \leq T_m - T_n\}}\}$ of $X_{(i,n)}$ at stage m consists of the possibly censored lifetime $Z^i_{n,m} = \min(X_{(i,n)}, T_m - T_n)$ and its indicator function

$$I_{\{X_{(i,n)} \leq T_m - T_n\}} = \begin{cases} 1 & \text{if } X_{(i,n)} \leq T_m - T_n \\ 0 & \text{otherwise.} \end{cases}$$

Example 3.1:
For the prototypical example, to model the censoring of the patients's lifetime we assume that $X_{i,n}$ follows the exponential distribution with unknown parameter θ and Y_n follows the exponential distribution with known parameter σ. The parameter $\frac{1}{\sigma}$ indicates the average inter-selection times between consecutive selections, and the number of selections occurring in a fixed time interval (0,T), which has a Poisson distribution with parameter σT. Then we can find that the $(n+1)^{st}$ selection time $T_{n+1}, n \geq 1$, has a gamma(n, σ) distribution. Conditional on θ, the possibly censored observation $Z^i_{n,n+1}$ has the distribution $P(Z^i_{n,n+1} \leq x|\theta) = 1 - e^{-(\theta+\sigma)x}$, and the indicator function $I_{\{X_{i,n} \leq Y_n\}}$ is a Bernoulli random variable with the parameter $\frac{\theta}{\theta+\sigma}$ as the probability of success.

Up to here all fundamental elements in the model have been defined, and in the later subsection we shall use them to apply the Bayesian approach.

3.3 The states and transition

Following the Bayesian approach, we can define a *dynamic stochastic* model, which can completely describe the randomness of the arms at the corresponding stages, and can indirectly describe the effectiveness of the treatments in clinical trials. It is very clear that the prior and posterior distributions play a key role in describing the state of our controlled processes. West and Harrison[41] have shown that a condition to describe the dynamic model that is the current state must depend on previous states and observations. Since a posterior distribution depends on the prior distribution, and the observations are in controlled stochastic processes, we are really interested to know whether this dependence is measurable in our model.

In order to achieve the dynamic model, we discuss *states* and *transition* in this subsection. They play the key role in the dynamic and randomness models. We denote the *information bank* by $(C_{i,n}, (Z^i_{(j),n}, j = 1, ..., C_{i,n}))$, $C_{i,n}$ is the patient on the unknow arm who are still living. It can completely describe the censored information on arm i at stage n. The *size* of the information bank $C_{i,n}$ is the total number of

censored observations on arm i at stage n. $Z^i_{(1),n}, ..., Z^i_{(C_{i,n}),n}$ are the *censored observations* on arm i at stage n, that represent the lifetimes of individuals who are still alive at a time before the decision on the next patient is made. If $C_{i,n} = 0$, then the information bank is *empty* and all lifetimes on arm i are uncensored at stage n. If there are k-armed selections, the posterior distribution is denote by G_n, and the random tuple

$$s_n = [G_n, (C_{i,n}, Z^i_{(j),n}, i = 1, ..., k, j = 1, ..., C_{i,n})]$$

completely describes the state of process with censored observations at stage n, and provides sufficient information for selecting an arm. Let S be the set of all such possible states s_n, and call it the *state space*. Following the state space and the action space, we can describe any transition of the process from one stage to another. Since arm i can be chosen randomly in the action space, and the censoring time y_n is the realized value of random variable Y_n, then the state s_{n+1} is random. This randomness can be described by the state transition probabilities. Wang[40] denotes a *transition probability* $Q_{n+1}(s_{n+1} \in H | s_n, i_n, y_n)$ in a conditional probability distribution as the probability that the state s_{n+1} at stage $n+1$ is in the measurable subset H of state space S, given the state s_n, selection i_n and value y_n. Depending on the different treatments, there are many cases for the new transition probability, and this is another randomness to be applied. Here we use an example to make it more clear.

Example 3.2:
For the prototypical example, consider one patient's case. Since our payoff distribution have a prior gamma distribution $g(\eta_n, \tau_n)$ with hazard rate θ, which has the loss of memory property, we can simply write the state of treatment $s_n = ((\eta_n, \tau_n), (c_n))$, where c_n is the total number of censored observations at stage n. Suppose the state at stage one is given as $s_1 = ((\eta, \tau), (0))$ and selected treatment is i_1, and the observation is y. Then the state at stage two is $s_2 = ((\eta_2, \tau_2), (c_2))$. Since the censoring times are continuous, we can denote a transition probability $Q_2(s_2 \in H | s_1, i_1, y)$ by a conditional probability of a measurable subset H of state space at stage 2. Depending on the treatment we selected initially, there are three cases for the new transition probability.

If the known treatment B at the first patient, i.e., $i_1 = 2$ then all the expected lifetimes λ are known, and the observation of its lifetime provides no information on treatment A, i.e., the information bank is empty and together with the initially number of censored observations $c_1 = 0$, i.e, data bank empty. That is to say, the state will be the same as the state at the first stage, even if there are some censoring times y.

On the other hand, if the unknown treatment A is selected initially and G_1 is the gamma (η, τ) distribution,

for any $t > 0$, the lifetime X under the state s_1 has a probability density function :

$$\begin{aligned} P(X \leq t) &= E(I_{\{X \leq t\}}) \\ &= E(E(I_{\{X \leq t\}})|\theta) \\ &= \int_0^\infty \int_0^t \theta e^{-\theta s} ds \frac{\eta^\tau}{\Gamma(\tau)} \theta^{\tau-1} e^{-\eta\theta} d\theta \\ &= \int_0^\infty (1 - e^{-\theta t}) \frac{\eta^\tau}{\Gamma(\tau)} \theta^{\tau-1} e^{-\eta\theta} d\theta \\ &= 1 - \left(\frac{\eta}{\eta + t}\right)^\tau. \end{aligned}$$

The marginal density function of X:

$$f(t) = \tau \eta^\tau \left(\frac{1}{\tau + t}\right)^{\tau+1}.$$

Let $Z = \min(X, y)$ and $\delta = I_{\{X \leq y\}}$, because of gamma distribution is conjugate distribution then the state s_2 is of the form:

$$s_2 = ((\eta + Z, \tau + \delta), (1 - \delta)).$$

If the censored patient still alive after censoring time y(i.e. $\delta = 0$) :

$$B_2(y) = \{((\eta + y, \tau), (1))\}$$

If the censored patient died during the censoring time y :

$$B_3(x, y) = \{((\eta + x', \tau + 1), (0)) : 0 \leq x' \leq x \leq y\},$$

where x' is the time between last arrival and death.

Then for any y and any $0 \leq x \leq y$, $B_2(y)$ and $B_3(x, y)$ are measurable subsets of S. We have

$$Q_2(s_2 \in B_2(y)|s_1, i_1 = 1, y) = P_1(X \geq y|s_1, i_1 = 1, y) = \left(\frac{\eta}{\eta + y}\right)^\tau.$$

And for any $0 \leq x \leq y$,

$$\begin{aligned} Q_2(s_2 \in B_3(x, y)|s_1, i_1 = 1, y) &= P_1(0 \leq X \leq x|s_1, i_1 = 1, y) \\ &= 1 - \left(\frac{\eta}{\eta + x}\right)^\tau. \end{aligned}$$

So s_2 is of mixed type, whose probability is $(\frac{\eta}{\eta+y})^\tau$ at one-point mass $((\eta + y, \tau), (1))$ and whose density is $\tau\eta^\tau(\frac{1}{\tau+t})^{\tau+1}$. on the interval $[0, y]$.

In this subsection, we have described the dynamics of a stochastic system. In the terminology of controlled stochastic processes, selecting an arm means taking an action. In our model, the action space is $A = \{1, ..., k\}$, so the selections are made among the k arms at various stages, and are specified by strategies. In the next subsection, we will discuss some details of strategies and their worths.

3.3.1 The strategies and their worths

In our model, a *strategy* is just a sequence of measurable functions, denote by $\pi_n : S \times A^{n-1} \to A$.

Define a *reward* $r(s,i,t)$, which is the expected value of the random lifetime $X_{i,n}$ from selecting arm i when state is s at time t. That is,

$$r(s,i,t) = E(X_{i,n}|G) = \int_{D^k} \int_0^\infty x dF_i(x) dG(F_1, ..., F_k),$$

where G is a posterior distribution on $(F_1, ..., F_k)$ that depends on t.

Following this structure we can define the worth of any strategy. Recall that F_0 is the distribution for the sequence of potential censoring times, and its nth convolution $F_0^{(n)}$ is the distribution of the $(n+1)$th selection time $T_{n+1}, n = 1, 2, ...$ The first selection occurs at time $T_1 = 0$, so we denote $F_0^{(0)}$ to be the discrete one-point distribution such that $F_0^{(0)}(x) = 1$ for all $x \geq 0$. For each $n = 1, 2, ...$ let $D_n : R^+ \to R^+$ be a discount function for the nth selection, such that $\sum_{n=1}^{\infty} \int_0^\infty D_n(t) dF_0^{(n-1)}(t) < \infty$. Consider that each strategy π specifies a sequence of selections of the arms $\{i_n, n = 1, 2, ...\}$ and develops a sequence of states $\{s_n, n = 1, 2, ...\}$. Therefore the strategy π generates a process $\{(s_n, i_n) : n = 1, 2, ...\}$ in the space $\bigtimes_0^\infty (S \times A)$. By Kolmogorov's existence theorem[22], we can define a probability measure P_π on $\bigtimes_0^\infty (S \times A \times R^+)$. So let E_π be the expectation taken with respect to P_π, and denote by $X_{(i_n,n)}$ the survival time generated by strategy π_i at stage n, which has the arrival time distribution $F_0^{(n-1)}(t)$, and denote $Z_n(t)$ the censored survival time distribution of n^{th} patient which depend on time t.

Now we can define the *finite horizon worth* with N selected patiens of the strategy π to be

$$\begin{aligned} W_N(s_1, \pi) =& E_\pi \left[\sum_{n=1}^{N} \int_0^\infty D_n(t) Z_n(t) dF_0^{(n-1)}(t) \right] \\ =& \sum_{n=1}^{N} \int_0^\infty D_n(t) [E_\pi Z_n(t)] dF_0^{(n-1)}(t); \end{aligned}$$

Example 3.3:
For the prototypical example, since the known treatment B gives known expected lifetime λ, then for any state s, and any time t, if $i = 2$ then $r(s,i,t) = \lambda$. On the other hand, since the effectiveness of treatment A

is unknown, s is of the form $((\eta,\tau),(c))$, then r is continuous in η,τ; and it is of the form

$$\begin{aligned} r(s,i,t) =& E(X_{(1,n)}|(\eta,\tau)) \\ =& \int_0^\infty \int_0^\infty x\theta e^{-\theta x} dx \frac{\eta^\tau}{\Gamma(\tau)} \theta^{\tau-1} e^{-\eta\theta} d\theta \\ =& \int_0^\infty \frac{1}{\theta} \frac{\eta^\tau}{\Gamma(\tau)} \theta^{\tau-1} e^{-\eta\theta} d\theta \\ =& \int_0^\infty \frac{\eta^\tau}{\Gamma(\tau)} \theta^{(\tau-1)-1} e^{-\eta\theta} d\theta \\ =& \frac{\eta}{\tau-1} \int_0^\infty \frac{\eta^{\tau-1}}{\Gamma(\tau-1)} \theta^{(\tau-1)-1} e^{-\eta\theta} d\theta \\ =& \frac{\eta}{\tau-1} \end{aligned}$$

.

Let us consider the worths of four non-adaptive strategies $\pi_1, \pi_2, \pi_3, \pi_4$, all with the initial state $s_1 = ((\eta,\tau),(0))$ and let $D = (a_1, a_2, ...)$ be a discount sequence.

Suppose π_1 chooses the treatment A at both stages, and stage two occurs after a delay time t. If the first patient has died at time $x \le t$, then the state at stage two is $s_2' = ((\eta + x, \tau+1),(0))$ and the the expected lifetime of second patient is $r(s_2', i_2 = 1, t) = \frac{\eta+x}{\tau}$. On the other hand, if the first patient is still alive at time t, then the state at stage two is $s_2'' = ((\eta+t,\tau),(1))$, and the the expected lifetime of second patient is $r(s_2'', i_2 = 1, t) = \frac{\eta+t}{\tau-1}$.

Therefore,

$$\begin{aligned} E_{\pi_1} Z_2(t) =& \int_0^t r(s_2', 1, t)\tau\eta^\tau \left(\frac{1}{\eta+x}\right)^{\tau+1} dx + \int_t^\infty r(s_2'', 1, t)\tau\eta^\tau \left(\frac{1}{\eta+x}\right)^{\tau+1} dx \\ =& \int_0^t \frac{\eta+x}{\tau}\tau\eta^\tau \left(\frac{1}{\eta+x}\right)^{\tau+1} dx + \int_t^\infty \frac{\eta+t}{\tau-1}\tau\eta^\tau \left(\frac{1}{\eta+x}\right)^{\tau+1} dx \\ =& \frac{\eta}{\tau-1} \end{aligned}$$

Finally, we get the worth:

$$\begin{aligned} W_2(s_1,\pi_1) =& a_1 \frac{\eta}{\tau-1} + a_2 \int_0^\infty \beta_2(t) \frac{\eta}{\tau-1} \sigma e^{-\sigma t} dt \\ =& a_1 \frac{\eta}{\tau-1} + a_2 \int_0^\infty e^{-at} \frac{\eta}{\tau-1} \sigma e^{-\sigma t} dt \\ =& a_1 \frac{\eta}{\tau-1} + a_2 \frac{\sigma}{\sigma+a} \frac{\eta}{\tau-1}. \end{aligned}$$

Suppose π_2 chooses treatment B at both states then the worth is

$$W_2(s_1,\pi_2) = a_1\lambda + a_2 \frac{\sigma}{\sigma+a}\lambda.$$

Suppose π_3 chooses treatment A at first stage, and treatment B at second stage then the worth is

$$W_2(s_1, \pi_3) = a_1 \frac{\eta}{\tau - 1} + a_2 \frac{\sigma}{\sigma + a} \lambda.$$

Suppose π_4 chooses treatment B at first stage, and treatment A at second stage then the worth is

$$W_2(s_1, \pi_4) = a_1 \lambda + a_2 \frac{\sigma}{\sigma + a} \frac{\eta}{\tau - 1}.$$

Up to here, we defined strategies and their worth equations. We are interested in the optimal strategies and the best worth equation of our model. This will be the topic of the following subsection.

3.4 Optimality equations and optimal strategies

In our model, the *optimality equations* are a sequence of functional equations which not only embody the principle of dynamic programming, and depict the recursive relations between the values of the bandit processes of various horizons, but also, and more importantly, lay a foundation for calculating the value and finding an optimal strategy in principle[40].

For any given $m = 1, 2, ...$, we suppose that stage m occurs at a given random time T, and the given state s_m has the posterior distribution G_m, which depends on t_m. Consider a strategy π that selects arm i at stage m and follows an optimal strategy from $m+1$, where the immediate expected reward is $D_m(t_m)r(s_m, i, t_m)$. Then we denote a value of optimal strategy:

$$\begin{aligned} V(m, s_m) &= \sup_{\pi} W(m, s_m, \pi) \\ &= \sup_{\pi} \sum_{i=1}^{k} \pi_m(\{i\}|s_m, i_1, ..., i_m - 1) \left[D_m(t_m) r(s_m, i, t_m) + \int_0^{\infty} \int_S V(m+1, s) Q_{m+1}(ds|s_m, i, t) dF_0(t) \right]. \end{aligned}$$

The right hand side without the *sup* is the worth of the strategy π from stage m on. The value of this strategy from stage $m+1$ is $V(m+1, s_{m+1})$, which depends on the state s_{m+1} at stage $m+1$ under this strategy. Since this strategy presents in principle the value of the bandit process from stage $m+1$ on, the value of the bandit process should be the *sup* over all such strategies.

Optimality equations play an essential role in any dynamic programming model. We hope to find an optimal strategy, but the arms are chosen in a random way so the set of all strategies is uncountable. How do we avoid the uncountable? The best method is try to restrict the randomized strategy to a non-randomized strategy. Due to Wang's *Theorem 2.4.1*, for any randomized strategy, there is a non-randomized strategy which dominates the randomized strategy. Hence, depending on our model suppose there exists an optimal strategy. It must be one which is non-randomized, then we can restrict ourselves to the set of all non-randomized strategies to find an optimal strategy. In order to prove the existence of an optimal strategy, we

reduced the optimality equations to simpler forms:

$$V(m, s_m) = \sum_{i=1}^{k} \left[D_m(t_m) r(s_m, i, t_m) + \int_0^{\infty} \int_S V(m+1, s) Q_{m+1}(ds|s_m, i, t) dF_0(t) \right].$$

We can use this simpler form, combined with *Theorem 2.5.1* in Berry and Fristedt[3] to give the following result. For any state $s_n \in S$, there must exist a non-randomized strategy π^* which is optimal, i.e., $V(s_n) = W(s_n, \pi^*)$. To prove this theorem, we can start with horizon models first, then test for the general infinite horizon processes. Wang provides a detailed proof in section 2.5 of his thesis[40], so we do not list here again.

For the motivated example, since an optimal strategy can be found in the set of all non-randomized strategies, so we put the focus on the worth $W(\pi)$ of a non-randomized strategy π. There are many questions of the properties of $W(\pi)$ should be discussed, first is whether $W(\pi)$ depends continuously on the parameters (η, τ); second is how to order the relations among parameters affects the values of $W(\pi)$. Before test these properties, we first examine the structure of the worth $W(\pi)$ on any strategy π.

3.4.1 The structure of the worth

Let us write $u = \frac{\eta}{\tau}$, then the hazard rate θ has expected value $E(\theta) = \frac{\tau}{\eta} = u^{-1}$. Suppose the unknown treatment has expected hazard u^{-1}, while the known treatment has constant hazard λ^{-1}. As shown at the past sections, the marginal density of lifetime T on the unknown treatment is $f(t) = \tau\eta^{\tau}(\frac{1}{\eta+\tau})^{\tau+1}$, the marginal distribution is $F(t) = 1 - (\frac{\eta}{\eta+\tau})^{\tau}$, and the expected value of lifetime is $\frac{\eta}{\tau-1}$. Suppose the trial starts at time 0, with Gamma (η, τ) prior distribution. And at any time t, we observed a total lifetime T which does not matter censored or uncensored, and a total number r of deaths on the unknown arm, then the posterior at time t is also a Gamma distribution with parameter $(\eta + T, \tau + r)$. Since the lifetimes follow an exponential distribution, which has the loss of memory property, the state at any stage n is of the form $s_n = ((\eta_n, \tau_n), (c_n))$. The observations at the next stage $n+1$ only depend on state s_n, and selection i_n made at stage n, as well as the potential censoring time y_n between the n^{th} and $(n+1)^{th}$ selections.

Suppose that at some stage, the state is $s = ((\eta, \tau), (c))$, let $T_1', ..., T_c'$ be the residual lifetimes for these c patients' censored lifetimes, and the next stage occurs at a random lag time Y later. Then the observations of these residual lifetimes at the next stage are $Z_i = \min(T_i', Y)$, and their indicator functions are $\rho_i = I_{\{T_i' < Y\}}$. Since there is only need to consider non-randomized strategies, consider π to be the set of all non-randomized strategies, and π_1 (π_2 respectively) to be the subset of π consisting of strategies which select the unknown

(known respectively) arm initially. Let N be any positive integer. Recall that

$$W_N(s_m, \pi) = E_\pi \left[\sum_{n=m}^{N} \int_0^\infty D_n(t) Z_n(t) dF_0^{(n-m)}(t) \right]$$

is the worth of the strategy π from stage m to N, $m \leq N$.

We now discuss the recursive equations for the worths of myopic strategies with various horizons. A *myopic strategy* is the one that maximizes the reward for the current patient. In our model, the myopic strategy is very important for simulations and can be considered as using the best treatment for the patient currently.

First suppose $\pi_2 \geq \pi_1$, i.e., the known treatment is better than the unknown treatment, and the known arm is selected initially. In addition, all future selections will also be the known arm if the first arm selected is the known treatment. For any given value l, there are $\begin{pmatrix} c \\ l \end{pmatrix}$ possible ways to select l deaths from c lifetimes. Since all lifetimes are independent and identically distributed , without loss of generality, let $t_1, ..., t_l$, be l uncensored observations. Then, $\int_y^\infty f(t)dt = q(y) = \left(\frac{\eta}{\eta+y}\right)^\tau$ for censored lifetimes, and $\min(T_i, y) = T_i$ for uncensored lifetimes T_i, we have

$$\begin{aligned} W_N(s, \pi_2) &= a_1 \frac{\eta}{\tau - 1} + E_{\pi_2}\left(W_N(2, s^{(2)}, \pi_2)\right) \\ &= a_1 \frac{\eta}{\tau - 1} + \sum_{l=0}^{c} \begin{pmatrix} c \\ l \end{pmatrix} \int_0^\infty \underbrace{\int_0^y \cdots \int_0^y}_{l} W_N(2, s^{(2)}, \pi_2) f(t_1) ... f(t_l) dt_1 ... dt_l [q(y)]^{c-l} \sigma e^{-(\sigma+a)y} dy, \end{aligned}$$

$$\begin{aligned} s^{(2)} &= \left((\eta^{(2)}, \tau^{(2)}), c^{(2)}\right), \\ \eta^{(2)} &= \eta + \sum_{i=1}^{c} Z_i = \eta + \sum_{i=1}^{l} t_i + (c-l)y, \\ \tau^{(2)} &= \tau + \sum_{i=1}^{c} \rho_i = \tau + l, \\ c^{(2)} &= c - \sum_{i=1}^{c} \rho_i = c - l. \end{aligned}$$

On the other hand, suppose $\pi_1 > \pi_2$, i.e., the unknown arm is selected initially. Let T_{c+1} be the random lifetime from this new selection. Then its observations at the next stage consists of $Z_{c+1} = \min(T_{c+1}, Y)$ and $\rho_{c+1} = I_{\{T_{c+1} \leq Y\}}$. The state at the next stage is random, and for the strategy π_1 we have

$$\begin{aligned} W_N(s, \pi_1) &= a_1 \frac{\eta}{\tau - 1} + E_{\pi_1}\left(W_N(2, s^{(1)}, \pi_1)\right) \\ &= a_1 \frac{\eta}{\tau - 1} + \sum_{l=0}^{c+1} \begin{pmatrix} c+1 \\ l \end{pmatrix} \int_0^\infty \underbrace{\int_0^y \cdots \int_0^y}_{l} W_N(2, s^{(1)}, \pi_1) f(t_1) ... f(t_l) dt_1 ... dt_l [q(y)]^{c+1-l} \sigma e^{-(\sigma+a)y} dy, \end{aligned}$$

where $q(y) = \left(\frac{\eta}{\eta+y}\right)^{\tau}$ is the probability of a censored observations at time y, and

$$
\begin{aligned}
s^{(1)} &= \left((\eta^{(1)}, \tau^{(1)}), c^{(1)}\right), \\
\eta^{(1)} &= \eta + \sum_{i=1}^{c+1} Z_i = \eta + \sum_{i=1}^{l} t_i + (c+1-l)y, \\
\tau^{(2)} &= \tau + \sum_{i=1}^{c+1} \rho_i = \tau + l, \\
c^{(1)} &= c + 1 - \sum_{i=1}^{c+1} \rho_i = c + 1 - l.
\end{aligned}
$$

Note that both $W_N(2, s^{(1)}, \pi_1)$ and $W_N(2, s^{(2)}, \pi_2)$ have horizon $N-1$, so these two equations reduce the horizon of the worth of any strategy. Now let's consider how to maximize the total discounted expected lifetimes when the horizon is 2. That is,

$$
\begin{aligned}
V(s, \lambda) &= sup_{\pi} E_{\pi} \left(a_1 Z_1 + a_2 \int_0^{\infty} e^{-ay} Z_2 \sigma e^{-\sigma y} dy \right) \\
&= sup_{\pi} E_{\pi} \left(a_1 Z_1 + \left(a_2 \int_0^{\infty} e^{-ay} \sigma e^{-\sigma y} dy \right) Z_2 \right).
\end{aligned}
$$

If we have immediate responses from the selections, the posterior distribution is the gamma $(\eta + T, \tau + 1)$ distribution. This is not affected by the particular selection time of the second selection. Therefore, the expected lifetime for the second selection is always subject to the discount

$$
a_2 \int_0^{\infty} \sigma e^{-(\sigma+a)y} dy = a_2 \frac{\sigma}{\sigma + a}.
$$

The worth of strategy π_1 is

$$
W(s, \lambda, \pi_1) = a_1 \frac{\eta}{\tau - 1} + a_2 \frac{\sigma}{\sigma + a} E\left(\frac{\eta + T}{\tau} \vee \lambda \right),
$$

where

$$
E\left(\frac{\eta + T}{\tau} \vee \lambda \right) = \lambda \left[1 - \left(\frac{\eta}{\eta + (a \vee 0)} \right)^{\tau} \right] + \frac{\eta}{\tau - 1} \left(\frac{\eta}{\eta + (a \vee 0)} \right)^{\tau - 1}.
$$

Similarly,

$$
W(s, \lambda, \pi_2) = a_1 \lambda + a_2 \frac{\sigma}{\sigma + a} \left(\frac{\eta}{\tau - 1} \vee \lambda \right).
$$

For comparing the difference between $W(s, \lambda, \pi^1)$ and $W(s, \lambda, \pi^2)$, there is a simple method. Define $\triangle(s, D) = W(s, D, \pi_1) - W(s, D, \pi_2)$ to be the advantage function of the unknown arm over the known arm. Then the sign of this function determines an optimal selection of the arm.

Let $u^+ = \max(u, 0)$ and $v^+ = \max(v, 0)$ where $u = \lambda\tau - \eta$ and $v = u - \lambda$. Then extending the equation we get

$$\begin{aligned}\triangle(s, \lambda) =& W(s, \lambda, \pi_1) - W(s, \lambda, \pi_2) \\ =& a_1 \left(\frac{\eta}{\tau - 1} - \lambda\right) + a_2 \frac{\sigma}{\sigma + a} \int_{m^+}^{n^+} \left(\frac{\eta}{\eta + y}\right)^{\tau} \left(\frac{\lambda\tau}{\eta + y} - 1\right) e^{-(\sigma+a)y} dy.\end{aligned}$$

Depending on this equation, we can get the following cases:

1. If $\eta \geq \lambda\tau$ then $u^+ = 0, v^+ = 0$, and

$$\triangle(s, \lambda) = a_1 \left(\frac{\eta}{\tau - 1} - \lambda\right) \geq 0.$$

Therefore, the unknown arm is optimal initially.

2. If $\lambda(\tau - 1) \leq \eta < \lambda\tau$, then $u^+ > 0$, $v^+ > 0$,

$$\triangle(s, \lambda) = a_1 \left(\frac{\eta}{\tau - 1} - \lambda\right) + a_2 \frac{\sigma}{\sigma + a} \int_{0}^{\lambda\tau - \eta} \left(\frac{\eta}{\eta + y}\right)^{\tau} \left[\frac{\lambda\tau}{\eta + y} - 1\right] e^{-(\sigma+a)y} dy \geq 0,$$

where $y \in [0, \lambda\tau - \eta]$ implies that $\eta + y \in [\eta, \lambda\tau]$, hence $\frac{\lambda\tau}{\eta+y} - 1 \geq 0$ for all $y \in [0, \lambda\tau - \eta]$. So the unknown arm is optimal initially.

3. If $0 < \eta < \lambda(\tau - 1)$, then $u^+ > 0, v^+ > 0$. Because $v = \lambda(\tau - 1) - \eta$, we have

$$\triangle(s, \lambda) = a_1 \left(\frac{\eta}{\tau - 1} - \lambda\right) + a_2 \frac{\sigma}{\sigma + a} \int_{\lambda(\tau-1)-\eta}^{\lambda\tau - \eta} \left(\frac{\eta}{\eta + y}\right)^{\tau} \left[\frac{\lambda\tau}{\eta + y} - 1\right] e^{-(\sigma+a)y} dy.$$

Let $\eta + y = t$, then we have

$$\triangle(s, \lambda) = a_1 \left(\frac{\eta}{\tau - 1} - \lambda\right) + a_2 \frac{\sigma}{\sigma + a} \eta^{\tau} e^{(\sigma+a)\eta} \int_{\lambda(\tau-1)}^{\lambda\tau} \left(\frac{1}{t}\right)^{\tau} \left(\frac{\lambda\tau}{t} - 1\right) e^{-(\sigma+a)t} dt.$$

It is clear that $\triangle(s, \lambda)$ is a continuous function of η, and for any $\eta \in (0, \lambda(\tau - 1))$ we have $\frac{\partial\triangle}{\partial\eta} > 0$ and $\frac{\partial\triangle}{\partial\tau} < 0$. Hence for any η it is an implicit function of τ for fixed λ, and also an implicit function of λ for fixed τ, there must exists an η^* such that $0 < \eta^* < \lambda(\tau - 1)$ and $\triangle(((\eta^*, \tau), 0), \lambda) = 0$. Moreover, if $\eta \geq \eta^*$ then the unknown arm is optimal; otherwise, the known arm is optimal. And when the other parameters are fixed, η^* is an increasing function of τ and λ. So the conclusion is $\exists$ an η^* such that the unknown treatment is optimal if and only $\lambda(\tau - 1) \geq \eta^*$.

So far we have discussed the solution to the motivated example with two selections, and we derived the critical equation for the case of an empty information bank, which determines the break-even values of η, τ and λ. In the next subsection we will try to find an optimal strategy π^* and find the the value of this optimal strategy.

3.4.2 The optimal strategies

Recall that if the initial state of the motivated example is $s = ((\eta, \tau), c)$, and the first selection of the arm is made at time 0, then the worth of a strategy π is

$$W(s, D, \pi) = E_\pi[\sum_{n=1}^{\infty} \int_0^{\infty} D_n(t) Z_n(t) dF_0^{(n-1)}(t)],$$

where $F_0^{(n-1)}$ is the distribution of the selection time of stage n, Z_n is the random lifetime generated by the strategy π at stage n, and $D_n = (a_1, a_2, ...)$ is the discrete discount sequence. We have also defined the value of the sequential selection model to be

$$V(s, D) = \sup_\pi W(s, D, \pi).$$

Our objective is to find an optimal strategy π^* from

$$V(s, D) = W(s, D, \pi^*).$$

.

It has been proved that we can restrict our search for an optimal strategy in the set Π of all non-randomized strategies. Actually, $\Pi = \Pi_1 \cup \Pi_2$, where $\Pi_1 \cap \Pi_2 = \emptyset$ and $\Pi_1(\Pi_2$ respectively) is the subset of Π consisting of those strategies which select the unknown (the known respectively) arm initially.

Let ϕ (φ respectively) be a strategy which selects initially the unknown (know) treatment, and then always selects the known (unknown) treatment at the following stages. Let $s_n^\phi = ((\eta_n^\phi, \tau_n^\phi), c_n^\phi)$ (s_n^φ respectively) be the state generated by the strategy ϕ (φ respectively) when the n^{th} patient is to be treated, and D is a geometric sequence then we can rewrite

$$V(s_n^\phi, D_n) = \triangle^+(s_n^\phi, D_n) + W(s_n^\phi, D_n, \pi_2),$$

and

$$V(s_n^\varphi, D_n) = \triangle^-(s_n^\varphi, D_n) + W(s_n^\varphi, D_n, \pi_1).$$

In order to show that under some conditions the delta function is increasing in η, and decreasing in τ and λ, we need to use the following equation[40] which is a continuous analogue of Theorem 2 in Eick[11].

For any initial state $s = ((\eta, \tau), c, D)$ and λ

$$\begin{aligned}\triangle((\eta,\tau),c,\lambda,D) = \left(a_1 - \sum_{j=2}^{n} a_j\right)\left(\frac{\eta}{\tau-1} - \lambda\right) &+ \sum_{j=1}^{n-1} \underbrace{\int_0^\infty \ldots \int_0^\infty}_{j} E\left[\triangle^+((\eta_j^\phi, \tau_j^\phi), c_j^\phi, \lambda, D_j)\right] h(t_1)\ldots h(t_j) dt_1 \ldots dt_j \\ &- \sum_{j=1}^{n-1} \underbrace{\int_0^\infty \ldots \int_0^\infty}_{j} E\left[\triangle^-((\eta_j^\varphi, \tau_j^\varphi), c_j^\varphi, \lambda, D_j)\right] h(t_1)\ldots h(t_j) dt_1 \ldots dt_j \\ &+ \underbrace{\int_0^\infty \ldots \int_0^\infty}_{n} E\left[V((\eta_n^\phi, \tau_n^\phi), c_n^\phi, \lambda, D_n)\right] h(t_1)\ldots h(t_n) dt_1 \ldots dt_n \\ &- \underbrace{\int_0^\infty \ldots \int_0^\infty}_{n} E[V((\eta_n^\varphi, \tau_n^\varphi), c_n^\varphi, \lambda, D_n)] h(t_1)\ldots h(t_n) dt_1 \ldots dt_n.\end{aligned}$$

The reader can find the proof of this in Wang[40].

To see that this delta function determines an optimal initial selection of the arm, Wang has already shown that $\triangle(s, D)$ is increasing in η, and decreasing in τ and λ for any D having finite horizon n, but we need to use a simulation to identify this conclusion again in next section.

In Wang's simulation, he used *Fortran* to simulate the experiments but there are some initialization problem. Fortran is a general-purpose, imperative programming language which is especially suited to numeric computation and scientific computing. But most of basic programs have not been included in the database, like the *generating random variables* which the error occurs in Wang's simulations.

Up to here, we have introduced all essential definitions and finished the theoretical framework. In the next chapter, we will try to use a simulation to confirm the properties of the function $\triangle$, and give some conclusions based on the simulation's result.

Chapter 4

Simulation and Results

4.1 Introduction

We are now very familiar with the prototypical example as it has been discussed since Chapter 1. In this chapter, we carry out simulations to analyze the prototypical example. We hope to gain insight into the structure and obtain some hints for suggesting optimal strategies. Suppose that there is a treatment to be allocated, and the initial state is $s = ((\eta, \tau), c)$. It has been shown that, when other parameters are fixed, there exists a unique break-even value η_N^* for η, if $\eta \geq \lambda(\tau - 1)$. In this case, exactly one of the unknown arm or known arm is optimal initially. Although this break-even value makes it is easy to define an optimal strategy, it is hard to find in practice. In the next subsection we discuss in detail what simulations have been done, and how they are implemented. In section 4.3 we proposed a strategy and compared it with myopic strategy, then present the results of the simulation and some discussion, including graphs and tables. In section 4.4 we discuss these results and give some general suggestions for choosing good strategies.

4.2 Simulation

In my simulation, I follow Wang's method but use a different software environment, and I have corrected an error on Wang's algorithm. Wang's program was designed in *Fortran*[12], where as my program is written in *R*[33]. My simulation also included a wider range of conditions than those found in Wang's thesis.

The inputs of the simulation are the following:

- the total number of selected patients N;
- the rate parameter of arrival time distribution σ;
- the constant expected survival time λ on the known treatment;
- the gamma prior distribution $g(\eta, \tau)$ for the hazard rate θ of the unknown treatment;
- the lower bound for searching optimal strategies b^*.

Generally, we observe the trial for the objective of maximizing the total non-discounted expected lifetimes from each patient's finite horizon. There are two sequences of random variates needed. The first sequence represents the patients' lifetime, which is exponential inter arrival times generated by the R function *rexp*. The second sequence represents the patients' lifetimes after being treated with either the known or unknown treatment. As the controlled stochastic process evolves, we observe the arrival time and the lifetimes on the treatment at any survival time. After we update the state of the controlled stochastic process, we make a selection of treatment, and generate new lifetimes on treatment. The lifetime is simulated using the true expected value θ, but the decision of which treatment to apply uses posterior distribution (η, τ). Selecting a treatment means taking an action from the action space $\{1, 2\}$, where if the unknown treatment is selected we take action 1, and take the action 2 for the known treatment. Suppose that at the time of the n^{th} selection, the updated state is $s_n = ((\eta_n, \tau_n), c_n)$. Under action 1 we generate the c_{n+1} exponential lifetime by the parameter θ, and receive a reward $\frac{\eta_n}{\tau_n - 1}$, while under action 2 we receive a constant reward λ. Then we generate the arrival time for the next patient, observe these lifetimes, and update the transition of the state at the next stage. The decision as to which action to take depends on the current state $s_n = ((\eta_n, \tau_n), c_n)$, the number $N - n$ of actions to be taken in the future, and the objective of maximizing the total non-discounted rewards from all the actions. Denote by $\pi_n^{(1)}$(and $\pi_n^{(2)}$ respectively) the strategy which selects the unknown (the known) treatment for the n^{th} patient, and then follows an optimal strategy for the remaining $N - n$ patients. Denote by $W_N(n, s_n, \pi_n^{(1)})$ and $W_N(n, s_n, \pi_n^{(2)})$ the worths of the $N-n+1$ patients under these two strategies $\pi_n^{(1)}$ and $\pi_n^{(2)}$, respectively. Then the maximum total expected lifetimes from all these N patients is

$$V_N(s_1) = \max\left(W_N(1, s_1, \pi_1^{(1)}), W_N(1, s_1, \pi_1^{(2)})\right), \tag{4.1}$$

which can be found in principle by the backward induction equations with the initial condition

$$V_N(s_N) = \max(\frac{\eta_N}{\tau_N - 1}, \lambda). \tag{4.2}$$

To specify an optimal selection of the treatment for the n^{th} patient, we need only to calculate the function

$$\triangle_N(n, s_n) = W_N(n, s_n, \pi_n^{(1)}) - W_N(n, s_n, \pi_n^{(2)}). \tag{4.3}$$

Select the unknown treatment optimally for the n^{th} patient if and only if

$$\triangle_N(n, s_n) \geq 0. \tag{4.4}$$

When $n = N$ we have

$$\triangle_N(s_N) = \triangle_N(N, s_N) = \frac{\eta_N}{\tau_N - 1} - \lambda. \tag{4.5}$$

In principle, by starting with this equation, and employing the backward induction method of dynamic programming, we could derive the function $\triangle_N(n, s_n)$ for any n. For the simulation code, we use a break-even value $\eta_n^* \in (0, \lambda(\tau_n - 1))$ such that $\triangle_N(n, s_n^*) = 0$, where $s_n^* = ((\eta_n^*, \tau_n), c_n)$. The unknown treatment is selected at the state $s_n = ((\eta_n, \tau_n), c_n)$ if and only if $\eta_n \geq \eta_n^*$. Therefore, we need only to consider those

strategies defined in the following way. At stage n, choose an $\overline{\eta}_n \in (0, \lambda(\tau_n - 1))$. If the updated state is $s_n = ((\eta_n, \tau_n), c_n)$, then select the unknown treatment if and only if $\eta_n \geq \overline{\eta}_n$. So any such point $\overline{\eta}_n$ defines a selection for the n^{th} patient, and we are looking for a point $\eta_n^* = \overline{\eta}_n$ which defines the optimal selection. Since $0 < \overline{\eta}_n < \lambda(\tau_n - 1)$, let $b_n = \frac{\overline{\eta}_n}{\lambda(\tau_n - 1)}$. Then $b_n \in (0, 1)$ and $\overline{\eta}_n = b_n \lambda(\tau_n - 1)$. In this case, it is equivalent to say that any point $b_n \in (0, 1)$ defines a selection for the n^{th} patient such that the unknown treatment is selected if and only if $\eta_n \geq b_n \lambda(\tau_n - 1)$. Hence, any such strategy π corresponds to a point $b = (b_N, b_{N-1}, ..., b_1) \in (0, 1)^N$, and the myopic strategy corresponds to the point $(1, 1, ..., 1) \in (0, 1)^N$. The worth $W(\pi)$ of the strategy π becomes a function $q(b)$ on the space $(0, 1)^N$, and

$$\sup_{\pi} W(\pi) = \sup_{b \in (0,1)^N} q(b). \tag{4.6}$$

So the search for an optimal strategy becomes finding a point $b^* = (b_N^*, ..., b_1^*) \in (0, 1)^N$ such that

$$q(b^*) = \sup_{b \in (0,1)^N} q(b). \tag{4.7}$$

We would like to emphasize that b_n^* may depend on λ, τ_n, c_n and other parameters from the example. Since b_n could be any number in the range (0,1), this may make the search of b_n^* difficult, especially true if n is large. Wang[40] found that $b^* = 0.97$ is a good lower bound to search for optimal strategies, so we directly applied this number to our simulations and tested its accuracy. If the amount of valid information gradually increased, the myopic strategy will eventually be in a dominant position, that is $b^* = (b_N^*, ..., b_1^*) = (1, 1, ..., 1)$. Typically a clinical trial for testing unknown treatment only chooses a small sample and determine whether to continue the experiment at half of the sample. Hence, we set the myopic strategy must be consistently better than the proposed strategy when the number of censored observations is over half the number of total patients. The censored observations in our simulation include the size of information bank c and the number of died patients τ. Because our simulation focuses on the difference between the proposed strategy and the myopic strategy, all the values of the result will be omitted if $\tau > \frac{N}{2}$ and $c > \frac{N}{2}$. Up to here, we described the main steps of our simulation and explained the effect of some functions in the program. More details and the codes can be found in the Appendix. In the next subsection, we will present and analyze some results which verify and extend Wang's simulations with a wider selection of parameters.

4.3 Results

Wang has discussed a number of simulations in his dissertation, including the *two patients case, three patients case* and *fourteen patients case*. All of them are used to test the parameters' properties. In particular, the *fourteen patients case* plays a key role in finding the properties of the function $\triangle$. Since there are three main parameters (N, σ, λ) in our simulation, and the function $\triangle$ is our main subject of this section, we will first use the new codes to simulate a *fourteen patients case*, and compare with Wang's results. In another subsection, we discuss the result for other value of the parameters.

4.3.1 Different selected patients N

In this subsection, we focus on the differences between Wang's simulation and my simulation. Some of the initial parameters for simulations have been specified. These include:

- the total number of selected patients $N = 14$;
- the rate parameter of arrival time distribution $\sigma = 0.5$;
- the constant expected lifetime $\lambda = 100$ on known treatment;
- a good lower bound $b^* = 0.97$ for searching optimal strategies.

We will propose a strategy and compare its performance with the myopic strategy by means of simulations with 10,000 replications. Based on these observations, we can clearly find the difference between two simulations and answer the question posed in Chapter 3: whether the function $\triangle(s, D)$ is decreasing as τ increases for constant N, σ, λ. All tables in this dissertation are mean values over 10,000 simulations. *Table 4.1*, *Table 4.2* and *Table 4.3* are copied from Wang's dissertation. In all tables, the parameter c represents the size of the information bank, and the parameter τ represents the number of dead patients. The results only show that the proposed strategy is consistently better than the myopic strategy.

Table 4.1: The value of proposed strategy based on Wang's codes, where N=14, $\sigma = 0.5$, $\lambda = 100$ and $b^* = 0.97$.

	τ					
c	2	3	4	5	6	7
0	1438.974	1414.461	1406.494	1403.006	1401.410	1400.919
1	1433.865	1413.283	1406.107	1403.012	1402.090	1401.830
2	1429.731	1412.311	1406.076	1403.661	1402.568	1402.536
3	1426.632	1411.380	1406.274	1403.893	1403.306	1403.137
4	1424.389	1410.683	1406.503	1404.260	1403.594	1403.613
5	1422.111	1410.012	1406.323	1404.608	1404.124	1404.093
6	1422.056	1410.458	1406.484	1405.111	1404.652	1404.591
7	1420.643	1410.503	1407.088	1405.581	1405.001	1404.769

Table 4.2: The value of myopic strategy based on Wang's codes, where N=14, $\sigma = 0.5$, $\lambda = 100$ and $b^* = 0.97$.

	τ					
c	2	3	4	5	6	7
0	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
1	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
2	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
3	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
4	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
5	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
6	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000
7	1400.000	1400.000	1400.000	1400.000	1400.000	1400.000

By observing *Table 4.2*, we find all values in Wang's myopic strategy are equal to 1400, which is unreasonable in a mean of simulation with 10,000 replications. There is only one way that can explain this case: all processes of myopic strategy chose the known treatment for the first patient then subsequent patients have opted for the same treatment. We trace back to Wang's program, and discovered that a constant variable was uninitialized which led the program to assign 0 to this constant variable and defaulted each patient's censored lifetime to 100.

Table 4.3: The advantage value of the proposed strategy over the myopic strategy based on Wang's codes, where $N = 14$, $\sigma = 0.5$, $\lambda = 100$ and $b^* = 0.97$.

	τ					
c	2	3	4	5	6	7
0	38.974	14.461	6.494	3.006	1.410	0.919
1	33.865	13.283	6.107	3.012	2.090	1.830
2	29.731	12.311	6.076	3.661	2.568	2.536
3	26.632	11.380	6.274	3.893	3.306	3.137
4	24.389	10.683	6.503	4.260	3.594	3.613
5	22.111	10.012	6.323	4.608	4.124	4.093
6	22.056	10.458	6.484	5.111	4.652	4.591
7	20.643	10.503	7.088	5.581	5.001	4.769

Table 4.4: The advantage value of the proposed strategy over the myopic strategy based on my codes, where $N = 14$, $\sigma = 0.5$, $\lambda = 100$ and $b^* = 0.97$.

	τ					
c	2	3	4	5	6	7
0	70.258	54.396	49.195	41.069	35.922	31.031
1	55.868	39.371	31.021	27.495	23.577	22.480
2	33.080	34.793	24.643	23.903	17.846	16.078
3	31.833	22.193	22.348	20.457	17.431	14.076
4	29.476	18.869	22.596	17.617	17.274	13.326
5	23.540	22.254	20.601	15.032	12.304	14.494
6	19.456	17.136	16.853	14.812	13.117	8.674
7	15.153	19.731	19.179	12.859	8.327	5.980

From *Table 4.3* and *Table 4.4*, we can intuitively find that the advantage values are decreasing as τ is increasing for the same number of c. By observing *Table 4.3*, the advantage value is decreasing as c is increasing only when τ <4, and the advantage value is increasing as c is increasing when τ >4. However, there is a big difference in *Table 4.4*: the advantage values are decreasing as c is increasing for all τ. In other words, *Table 4.3* strongly verifies the conclusion we got from Chapter 3. In addition, we find that the last value of the first row in *Table 4.3* is an approximate threshold of myopic strategy over proposed strategy, i.e., there must exist a break-even value in the process when $c = 0$ and τ increase from 7 to 8. However, this case does not happen in *Table 4.4*. We can find that the minimum number 5.98 appeared when $c = 7$ and $\tau = 7$, but it still does not meet the approximate threshold of the break-even value. There are many reasons that can cause this case to happen, but the main reason is that the $b^* = 0.97$ is not suitable for our codes. Hence, we should find an exact b^* before testing other parameters.

We tried many kinds of situations, however there is no b^* that can meet a good lower bound for searching optimal strategies based on $N = 14, \sigma = 0.5$, and $\lambda = 100$. Hence, we had to change some values of the parameters to find an exact b^*. Finally, we found $b^* = 0.93$ is a good lower bound for searching optimal strategies based on $N = 28, \sigma = 0.5, \lambda = 100$. Hence, the total number of patients $N = 14$ is a good condition for Wang's simulation, but it is not good enough for our simulation. So in the following, we will use $b^* = 0.93$ for our simulations and increase the total number of patients N from 14 to 28. Since we focus on the interaction between proposed strategy and myopic strategy, in the following tables we just list the advantage values of the proposed strategy over the myopic strategy.

Table 4.5: The advantage values of the proposed strategy over the myopic strategy based on my codes, where $N = 28, \sigma = 0.5, \lambda = 100$ and $b^* = 0.93$.

	τ												
c	2	3	4	5	6	7	8	9	10	11	12	13	14
0	95.109	92.588	87.348	73.969	66.240	61.842	57.814	53.570	49.469	33.479	22.954	12.609	7.751
1	85.891	66.234	61.350	54.120	48.074	42.391	38.341	38.689	33.123	21.810	18.428	10.068	6.141
2	62.073	57.062	45.747	44.715	37.918	36.898	32.005	31.346	27.227	18.805	15.927	13.869	5.754
3	53.231	42.223	40.695	41.994	32.009	32.888	31.973	28.061	25.454	15.839	16.639	12.942	4.922
4	49.634	37.839	36.414	30.396	32.222	30.922	25.141	23.024	24.220	14.098	11.078	9.117	4.798
5	46.237	42.325	33.198	26.533	30.567	24.888	22.557	22.120	22.491	16.542	10.403	8.559	4.492
6	38.237	34.704	30.113	30.371	28.024	25.535	20.499	18.501	17.881	12.304	13.096	11.772	3.935
7	42.844	35.985	27.403	25.031	16.889	25.206	22.573	18.202	17.441	16.252	11.168	10.919	3.579
8	31.050	27.451	26.681	23.435	21.857	22.225	21.978	20.237	16.659	14.165	12.583	10.876	3.237
9	29.671	23.425	22.424	22.684	19.537	20.295	18.555	17.261	16.230	15.189	11.228	8.168	2.633
10	24.020	22.964	21.203	20.973	21.618	18.422	16.724	17.684	10.011	9.739	8.864	9.864	3.033
11	22.945	18.259	15.491	17.688	16.924	15.891	14.249	10.831	13.388	10.199	9.188	7.336	2.460
12	18.608	14.692	14.456	13.980	15.102	11.524	12.693	10.370	13.658	12.230	8.633	5.491	1.945
13	16.022	13.323	13.416	12.625	10.306	10.930	9.850	9.630	10.277	10.925	7.397	4.559	1.485
14	14.229	12.846	12.557	10.796	9.392	9.081	7.642	5.354	4.547	4.696	2.256	2.818	0.487

By observing *Table 4.5*, we see that the values are decreasing as τ increases, and the proposed strategy decreased faster than the myopic strategy. This indicates that if we have the same censored patients, more patients died cause fewer information collected and lowering the effect of treatment. On the other hand, the values increased as c increased in both tables, and the proposed strategy increased slower than the myopic strategy. This indicates that if we have the same number of death patients, then the effect of treatment depends on the size of the censored patients. As more information is collected, the myopic strategy becomes optimal if the information is useful. However, the limitations of collected information, along with the patients be treated the effective information collection rate will gradually decrease then the efficiency of two strategies will both increase but the rate gradually decreases. Since the myopic strategy always picks up the better treatment, its advantage should increase faster than the proposed strategy after collecting of sufficient information. For more details, we draw graphs that offer an intuitive observation.

There are two graphs in *Figure 4.1*, both of which are two-dimensional. The first graph represents the advantage values of the proposed strategy over the myopic strategy with different number of dead patients, and the second graph represents the advantage values with different number of censored patients. By observing the first graph, we find that the advantage values are decreasing as τ increases, the range of advantage value decreased from over 85 to slightly larger than 0. The mean of the advantage values decreasing marginally, and it decreased from over 50 to slightly larger than 5. On the other hand, we can find the advantage values are decreasing as c increases in the second graph. The range of advantage values decreased from over 95 to slightly larger than 0, and the mean of advantage values have a larger range of diminishing, as it decreased from over 55 to slightly less than 10. In addition, we also draw some three-dimensional graphs to show the overall trends and the relationship between τ and c.

Figure 4.1: Boxplot for the advantage of the proposed strategy over the myopic strategy, where $N = 28, \sigma = 0.5, \lambda = 100$ and $b^* = 0.93$.

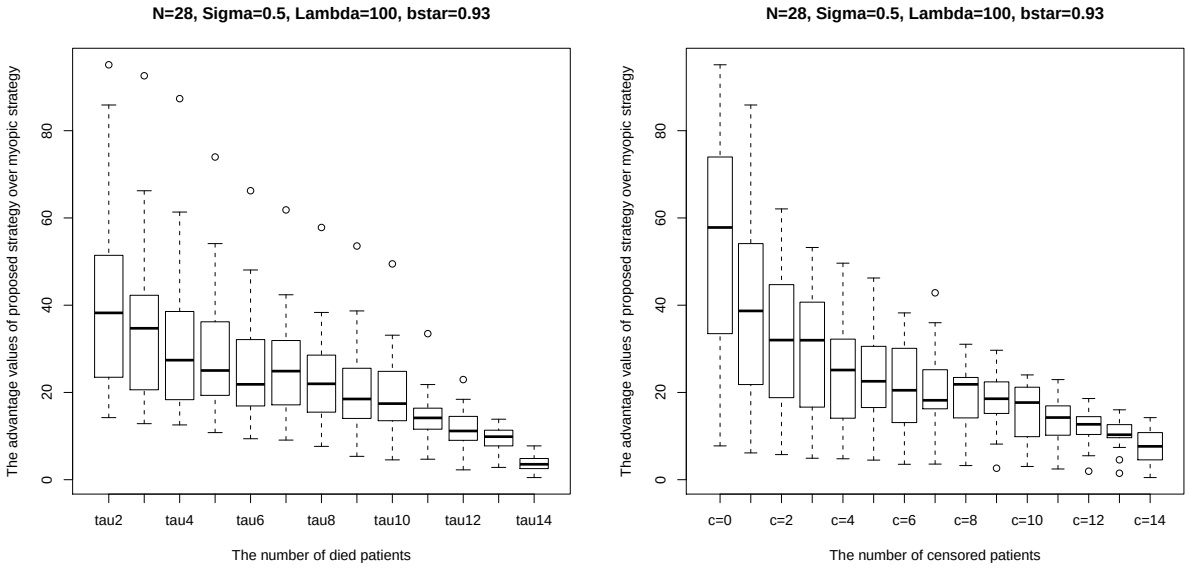
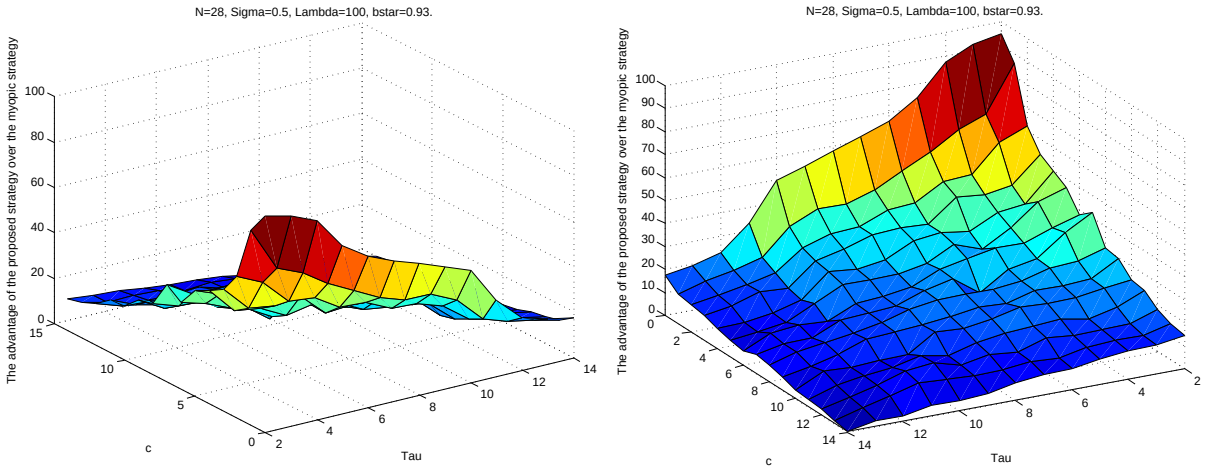


Figure 4.2: Three-dimensional graphs for the advantage of the proposed strategy over the myopic strategy, where $N = 28, \sigma = 0.5, \lambda = 100$ and $b^* = 0.93$

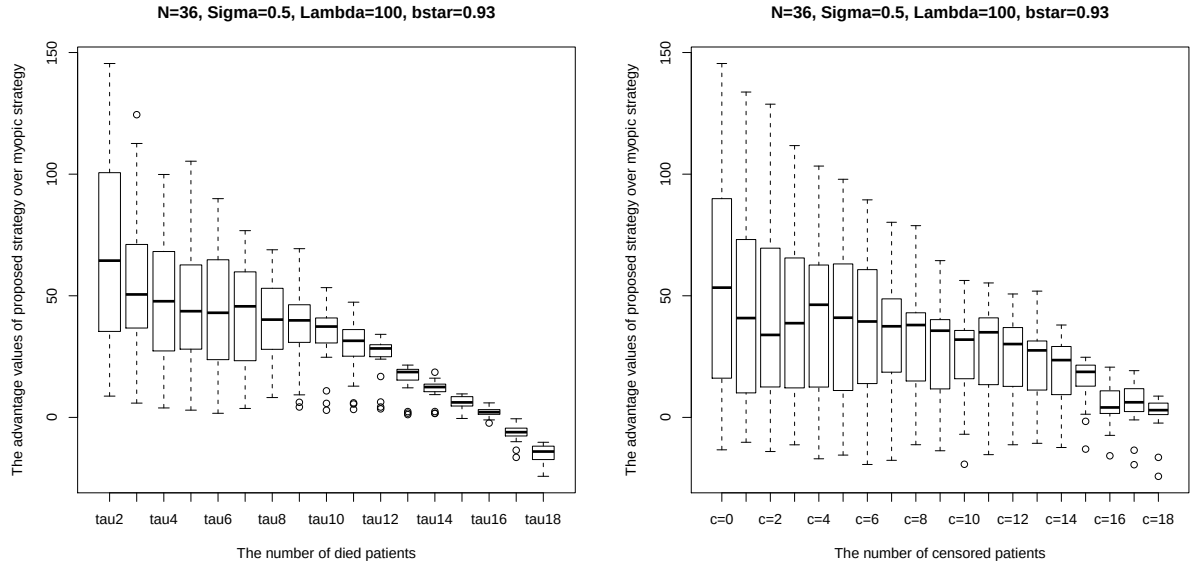


.

There are also two three-dimensional graphs in *Figure 4.2*. In order to more clearly observe, the second graph is the first graph after rotating 180 degrees. By observing *Figure 4.2*, we find the advantage values formed an inclined surface. This indicates that as c and τ moved towards 14, the advantage of the proposed strategy over the myopic strategy decreased in general.

We are also interested in testing the changes if we increase the total number of patients. In the follow up simulation, all the initial parameters are unchanged but N is increased to 36. By observing the left graph of *Figure 4.3*, we intuitively find that the mean of advantage values decreased very sharply. The range of advantage values reduced from slightly less than 150 to over -20. And all advantage values are negative when $\tau > 16$, which means the myopic strategy completely dominates the proposed strategy if $\tau > 16$. On the other hand, by observing the right graph of *Figure 4.3*. We find the mean advantage values decreases very slowly, and the break-even values appeared at each number c. Comparing with *Figure 4.2*, we can say that after N increased, the range of advantage values became wider and the myopic strategy completely dominated the proposed strategy before half the number of total patients. Hence, if the number N is increased, more information is collected and it is more suitable for the myopic strategy.

Figure 4.3: Boxplot for the advantage of the proposed strategy over the myopic strategy, where $N = 36, \sigma = 0.5, \lambda = 100$ and $b^* = 0.93$.



Through the analysis of selected patients N, we can say the myopic strategy is asymptotically optimal for a large number of total patients. In addition, we also verified the changing trends in the function $\triangle$ and verified a good lower bound $b^* = 0.93$ for searching for the optimal strategy. In the next subsection, we will see that detect the change of our simulations depends on different censoring times.

4.3.2 Different censoring time

In our model, there is an exponential inter-selection time which cumulative distribution function $H(t) = 1 - e^{-t\sigma}$, where $t \geq 0$. The expected value of inter-selection time indicates the average censoring time of the clinical trials, and the rate parameter σ plays a key role in arrival lifetime distribution. Hence, we can use the different σ to achieve the changes in simulation which depends on different censoring times. In this section, we will gradually reduce the value of σ from 0.5 to 0.1 which means the censoring time increases from 2 days to 10 days. Since the case of $\sigma = 0.5$ has been listed in *Table 4.5*, we just list the values for $\sigma = 0.2$ and $\sigma = 0.1$ in this section. In order to test the effect of different σ, some of the initial parameters for simulations have been specified. These include:

- the total number of selected patients $N = 28$;
- the constant expected survival time $\lambda = 100$ on known treatment;
- a good lower bound $b^* = 0.93$ for searching optimal strategies.

Table 4.6: The advantage values of the proposed strategy over the myopic strategy based on my codes, where $N = 28, \sigma = 0.2, \lambda = 100$ and $b^* = 0.93$.

	τ												
c	2	3	4	5	6	7	8	9	10	11	12	13	14
0	91.068	88.965	82.912	79.784	78.102	68.021	61.187	52.059	43.085	33.853	21.316	11.920	9.569
1	89.317	77.481	75.966	72.162	70.348	63.011	50.152	47.619	35.882	29.711	18.858	10.677	6.062
2	80.587	67.224	64.919	68.482	62.842	60.723	53.993	45.265	32.234	25.473	11.498	4.072	1.056
3	70.091	63.266	59.991	59.613	57.719	53.271	46.080	38.787	30.120	18.722	10.961	2.116	0.983
4	64.961	56.562	48.021	49.147	48.675	45.320	38.877	31.172	28.803	21.871	7.377	1.198	-0.361
5	50.997	52.663	44.258	43.107	42.523	41.473	31.360	28.355	25.564	23.508	6.962	0.233	-1.564
6	49.095	45.904	39.330	38.299	35.038	36.230	30.691	23.100	22.268	20.186	4.210	-1.761	-3.428
7	39.651	38.953	32.929	29.357	22.448	21.737	27.801	19.428	19.166	16.964	3.423	-3.249	-5.347
8	47.408	33.748	28.884	23.759	18.097	22.912	23.716	12.397	17.379	10.123	2.351	-3.843	-8.203
9	38.088	29.312	22.882	18.880	15.175	14.629	12.428	8.090	5.978	4.480	2.109	-5.408	-10.993
10	29.973	28.858	18.557	13.474	11.645	10.316	9.318	6.702	3.772	2.709	0.359	-6.560	-13.045
11	24.403	21.941	14.223	10.879	9.455	7.194	6.410	5.971	1.570	0.170	-4.952	-8.145	-15.435
12	18.274	13.078	11.296	9.541	6.934	4.943	3.862	1.269	0.973	-5.784	-9.875	-11.804	-19.224
13	10.301	9.519	7.780	4.933	2.707	2.145	1.245	0.793	-1.997	-7.842	-10.878	-13.079	-22.865
14	6.816	5.056	3.714	0.761	1.936	1.024	0.768	-2.171	-5.984	-9.036	-12.828	-16.830	-27.341

By observing *Table 4.6*, we find that the overall trends are the same as in *Table 4.5*. That is, the advantage values are decreasing as τ and c increase. We also find that the range of advantage values is increasing and lots of negative numbers appear. These illustrate that decreasing the number of σ can make the myopic strategy dominate the proposed strategy faster. The reason is that when σ decreased from 0.5 to 0.2, the censoring time increased from 2 days to 5 days. The increased censoring time must cause more information to be collected. This information is embodied in the number of patient deaths. For example, if a patient lived for 3 days after treatment, then they should be considered as dead if the censoring time is 5 days. On the other hand, they may be considered as alive if the censoring time is 2 days. In another words, σ just determines the time of gathering information in proposed strategies especially if the inverting for very little information at initial phase.

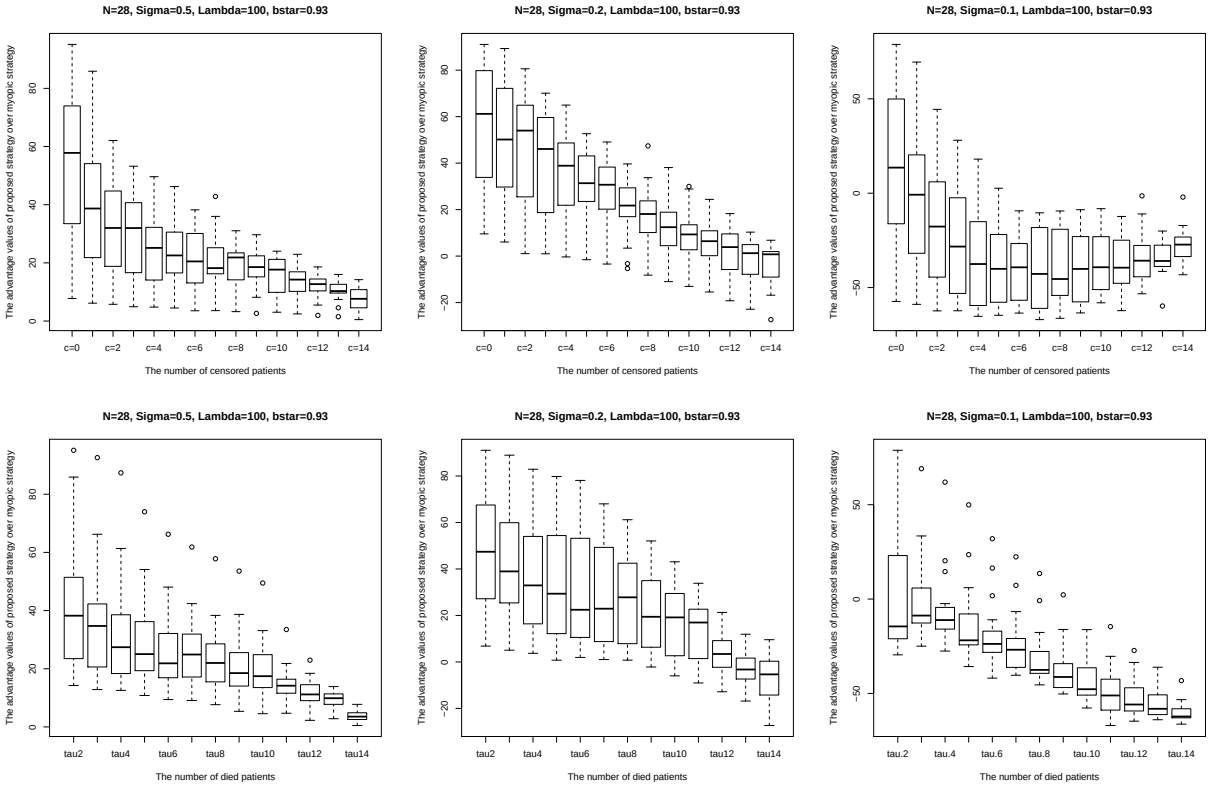
Table 4.7: The advantage values of the proposed strategy over the myopic strategy based on my codes, where $N = 28, \sigma = 0.1, \lambda = 100$ and $b^* = 0.93$.

	τ												
c	2	3	4	5	6	7	8	9	10	11	12	13	14
0	78.838	69.077	61.969	49.924	31.980	22.380	13.558	2.256	-16.187	-14.609	-27.239	-46.512	-57.345
1	69.491	33.446	20.310	23.541	16.422	7.266	-0.794	-16.091	-22.658	-31.837	-49.621	-50.942	-58.892
2	44.433	23.432	14.559	6.033	1.763	-6.636	-17.682	-30.617	-34.232	-44.463	-50.392	-58.138	-62.348
3	28.059	6.381	-2.407	-0.685	-10.975	-16.966	-28.277	-39.490	-48.174	-53.101	-57.429	-60.479	-62.293
4	18.051	5.331	-6.438	-15.077	-29.299	-25.773	-37.527	-41.846	-50.714	-59.887	-63.529	-59.498	-65.234
5	2.634	-12.143	-17.569	-21.884	-23.757	-37.798	-40.089	-41.289	-52.532	-57.783	-64.697	-63.940	-62.324
6	-18.491	-9.333	-11.118	-26.666	-28.411	-38.930	-39.264	-44.917	-49.862	-60.375	-56.717	-61.958	-63.519
7	-10.394	-14.753	-14.714	-18.189	-41.904	-34.588	-42.771	-49.188	-57.704	-66.984	-61.025	-63.665	-64.253
8	-14.493	-9.401	-13.185	-19.099	-25.656	-40.365	-45.464	-47.084	-51.062	-54.166	-63.455	-60.173	-66.274
9	-20.252	-8.743	-19.919	-22.983	-34.474	-40.127	-38.669	-50.222	-52.782	-60.618	-57.564	-63.428	-61.921
10	-17.495	-24.951	-8.188	-21.857	-23.026	-29.448	-39.227	-50.278	-47.758	-51.074	-55.868	-58.074	-53.298
11	-24.897	-13.228	-12.309	-24.676	-28.071	-26.845	-39.460	-46.679	-47.734	-46.676	-54.415	-57.645	-62.231
12	-25.719	-1.398	-10.972	-35.755	-27.668	-30.655	-36.089	-33.325	-39.515	-44.282	-44.393	-50.544	-53.283
13	-29.558	-20.048	-27.538	-34.159	-23.633	-26.420	-35.912	-38.819	-38.594	-40.606	-37.666	-41.450	-59.773
14	-21.799	-2.069	-17.105	-23.860	-23.216	-24.825	-27.234	-35.110	-29.207	-30.337	-33.563	-36.125	-43.205

In order to more clearly verify the characteristic of σ, we further decrease its value to 0.1. Although we just take it from 0.2 down to 0.1, the censoring time has increased by 5 days. By observing *Table 4.7*, we find more negative numbers which indicate the advantage of proposed strategies disappears faster than before. Also different from the past, the advantage values sharply decline when c is smaller than 5 and τ is larger than 7. Relative to *Table 4.6*, the difference between the two strategies is almost the same at the initial phase, but a large gap appears at the finial phase. As c increased, the advantage values also sharply decline after a short stabilization. Compared with *Table 4.4*, the σ decreased from 0.5 to 0.1 caused the censoring time to increase from 2 days to 10 days. This is a big variable change in clinical trials, since there are more days of censoring, and more patients are considered dead. It is clear that 8 days delay causes a big difference. There is an overall reduction of advantage values in *Table 4.7*. This means that $\sigma = 0.1$ not only reduced the gap of the two strategies at the initial phase but also reduced the time of the myopic strategy over the proposed strategy.

Figure 4.4 shows the changes when σ decreases from 0.5 to 0.1. By observing the advantage values as they change with c (upper row of *Figure 4.4*), we find that the mean advantage values change from sharp to flat. When $\sigma = 0.5$ all advantage values are positive and their mean converge to 20, so the proposed strategy is completely in a dominant position. When $\sigma = 0.2$ most of the advantage values are positive but a few of them are negative, which indicates the myopic strategy is gradually better than the proposed strategy. When $\sigma = 0.1$, we find that most of the advantage values are negative and their mean converge to -20, so the myopic strategy is in a dominant position. On the other hand, by observing the advantage values change with τ (lower row of *Figure 4.4*), we find that the mean of advantage values are always declining. In particular, all of the mean of advantage values are negative when $\sigma = 0.1$.

Figure 4.4: Boxplot for the advantage value of the proposed strategy over the myopic strategy with different number of σ



Through the above analysis, we found the role of σ in our simulations. With the reduction of the variable σ, the censoring time gradually increased. Longer censoring time should cause the number of dead patients increase. Hence, we think a smaller σ means more information was collected in the unknown treatment. The valid information more saturated cause the efficiency of proposed strategies to decrease faster, but the myopic strategy always keeps the most efficient solution. Hence, the advantage value of the proposed strategy over the myopic strategy should decrease faster. In the next subsection, we will detect the last key variable λ which is the expected maximum lifetime in the known treatment.

4.3.3 Different known expect lifetime

In our prototypical example, λ represents the expectation of survival time for the known treatment. By analyzing the advantage values of the proposed strategy over the myopic strategy based on different choices of λ, we can detect its role in our model and the influence on both strategies.

In this subsection, we applied three different choices of λ to our simulation. Since the case of $\lambda = 100$ has been listed in *Table 4.5*, we just list the tables for $\lambda = 90$ and $\lambda = 80$ here. In order to test the effect on the different choices of λ, some of the initial parameters for the simulation have been specified. These include:

- the total number of selected patients $N = 28$;
- the rate parameter of arrival time distribution $\sigma = 0.5$;
- a good lower bound $b^* = 0.93$ for searching optimal strategies.

Table 4.8: The advantage values of the proposed strategy over the myopic strategy based on my codes, where $N = 28, \sigma = 0.5, \lambda = 90$ and $b^* = 0.93$.

							τ						
c	2	3	4	5	6	7	8	9	10	11	12	13	14
0	62.822	52.067	46.638	37.497	32.582	28.993	20.955	13.182	8.382	4.992	2.526	-3.317	-4.059
1	45.111	36.755	30.769	24.327	20.918	18.148	12.327	8.830	3.747	1.755	-1.379	-2.427	-2.254
2	32.371	29.767	22.065	20.559	16.849	16.281	10.262	7.918	4.127	-0.263	-1.046	-1.427	-0.970
3	30.889	22.124	20.224	19.027	15.623	14.192	11.154	5.594	1.756	-1.351	-1.237	-2.008	-3.466
4	33.319	24.323	17.534	14.337	10.802	13.973	9.860	5.508	2.239	-3.220	-4.248	-4.318	-5.687
5	14.386	20.468	13.775	15.579	11.822	8.375	4.713	2.069	0.339	-2.856	-5.199	-4.971	-7.449
6	18.812	18.462	14.334	9.928	8.432	5.801	4.266	1.852	-2.924	-4.819	-5.274	-5.552	-8.800
7	15.885	15.988	10.786	10.687	10.106	8.413	6.870	1.773	-3.498	-5.331	-5.679	-8.592	-8.307
8	16.875	7.407	10.395	11.176	10.523	9.607	6.311	-2.750	-4.491	-6.313	-5.166	-7.137	-10.666
9	13.711	13.928	8.733	7.242	3.460	1.490	-3.547	-4.730	-6.860	-4.849	-9.722	-11.623	-12.023
10	13.690	9.762	11.018	7.952	7.914	5.181	4.945	0.305	-3.604	-3.777	-5.428	-9.177	-11.931
11	11.609	6.953	4.915	6.340	7.461	4.412	2.821	-4.213	-7.224	-4.455	-8.474	-11.146	-13.367
12	14.943	7.132	7.623	4.660	2.660	0.541	-2.256	-4.234	-9.934	-7.965	-10.702	-13.822	-11.179
13	10.496	8.304	5.165	1.523	-2.511	-3.521	-4.976	-7.830	-5.999	-9.626	-13.996	-10.737	-12.072
14	8.324	4.031	0.479	-1.892	-4.324	-8.838	-6.425	-11.333	-8.421	-10.923	-15.855	-14.311	-16.645

By observing *Table 4.8*, we find that a slight reduction of advantage values appears at the initial phase, which means a reduced λ causes the gap between the two strategies to decrease at the initial phase. The reason is that a smaller λ should lead more strategies to choose the unknown treatment at the initial phase. Hence, the gap must be decreased by more patients that received the same treatment. In addition, the advantage values are decreasing very slowly as τ and c increase, which means the proposed strategy and myopic strategy both have a steady upward trend but the myopic strategy is faster than the proposed strategy. Compared with *Table 4.5*, we find that the range of advantage values increased and more values are becoming negative. Hence, if we reduce the number λ, it causes the myopic strategy to completely dominate the proposed strategy earlier. In order to clearly verify the characteristic of λ, we further decrease its value to 80.

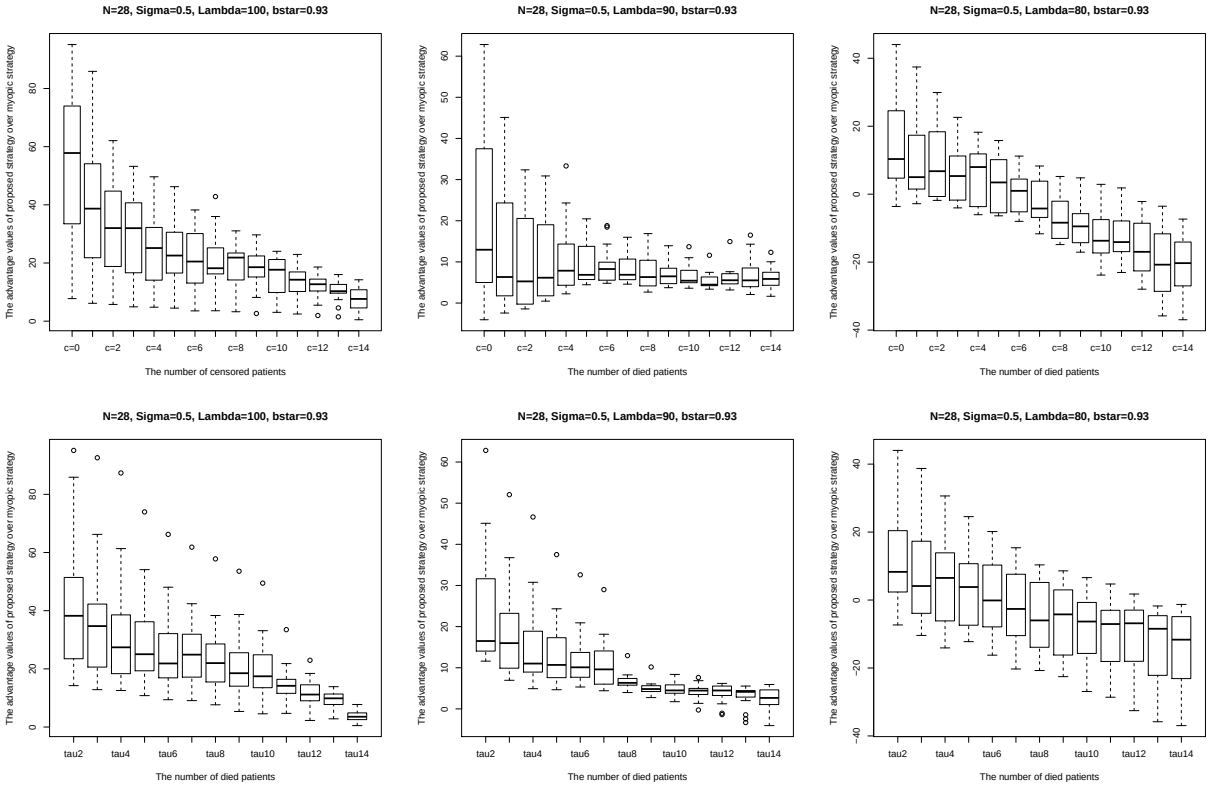
Table 4.9: The advantage values of the proposed strategy over the myopic strategy based on my codes, where $N = 28, \sigma = 0.5, \lambda = 80$ and $b^* = 0.93$.

							τ						
c	2	3	4	5	6	7	8	9	10	11	12	13	14
0	44.037	38.726	30.630	24.547	20.157	15.385	10.341	8.588	6.599	4.717	1.767	-2.965	-3.624
1	37.435	28.768	22.534	17.350	11.713	7.711	5.013	3.578	2.550	1.501	-2.150	-1.750	-2.802
2	29.912	23.274	21.647	18.366	11.715	8.486	6.765	4.780	3.812	-0.685	-1.124	-1.828	-1.291
3	22.605	19.149	15.432	11.238	8.877	7.469	5.337	2.421	0.231	-2.398	-1.747	-3.282	-4.036
4	18.227	15.460	12.348	9.869	11.853	9.364	7.981	3.556	-1.649	-3.657	-3.793	-5.921	-6.040
5	15.786	14.185	11.152	10.155	8.709	6.661	3.452	0.554	-2.883	-5.475	-6.372	-6.234	-5.729
6	11.211	9.082	7.365	4.439	3.091	1.028	0.975	-3.466	-6.030	-4.692	-5.196	-7.787	-8.005
7	8.289	4.123	6.504	3.827	-0.121	-2.625	-5.998	-4.222	-6.329	-7.067	-6.845	-8.459	-11.662
8	5.214	2.377	0.010	-2.071	-3.946	-8.394	-7.164	-10.582	-12.952	-14.837	-13.675	-13.018	-14.186
9	4.806	1.828	-4.951	-6.264	-5.732	-6.733	-9.477	-11.737	-13.792	-14.276	-16.431	-15.908	-17.067
10	2.893	-3.121	-5.838	-11.974	-7.487	-10.157	-13.707	-15.473	-16.593	-17.306	-19.635	-23.810	-23.198
11	1.840	-6.740	-7.882	-5.404	-8.307	-10.777	-14.097	-16.910	-14.841	-18.857	-16.346	-20.517	-23.0431
12	-2.168	-4.704	-6.452	-8.556	-13.874	-14.491	-16.969	-18.021	-20.566	-22.588	-24.938	-24.224	-27.965
13	-3.545	-5.576	-8.721	-11.666	-15.319	-17.532	-20.748	-22.547	-26.912	-28.601	-32.557	-35.819	-35.586
14	-7.332	-10.417	-14.090	-12.245	-16.224	-20.294	-19.201	-21.256	-24.332	-26.987	-29.684	-32.634	-36.957

By observing *Table 4.9*, we find that the advantage values are greatly reduced at the initial phase, and the lower bound range becomes more negative. The overall trend of advantage values decreased very slowly as τ and c increased. Compared with *Table 4.8*, we find that more advantage values become negative, and do so sooner. This indicates that the myopic strategy dominates the proposed strategy earlier. In order to provide a visual observation, we graph the advantage values for different choices of λ.

Figure 4.5 shows the changes when λ gradually decreases from 100 to 90, then to 80. By observing the advantage values as they change with c (upper row of *Figure 4.5*), we find that the upper bound range of advantage values gradually decreased and the lower bound range gradually increased. When λ decreased from 100 to 90 the mean of the advantage values changed from sharp to flat and converged to 10. Most of the advantage values are positive, with only a few negative numbers appearing when $c < 3$. When we decreased λ from 90 to 80, the mean of the advantage values changed back from flat to sharp and converged to -20. It is obvious that the proposed strategy is completely dominant since most of the advantage values are negative. On the other hand, by observing the advantage values as they change with τ (lower row of *Figure 4.4*), we find that the mean of the advantage values are always declining and converge to 5 when λ changes from 100 to 90. When we decreased λ from 90 to 80, most of the mean of advantage values are negative and converge to -5. Hence, we can say a smaller λ causes more advantage values to be reduced at the initial phase, and the overall trend is that the graphs always decreases at a constant rate as τ and c increase.

Figure 4.5: Boxplot for the advantage value of the proposed strategy over the myopic strategy with different number of λ



4.4 Summary

Motivated by clinical trial problems, this dissertation studies sequential selection with censored lifetime. The objective is to maximize some combination of total expected rewards from all selections. Following the Bayesian approach and employing the technique of dynamic programming, these sequential selection processes are formulated as general controlled stochastic processes. Some of the results for sequential selection with immediate responses are generalized to the sequential selection with censored lifetime, and most of these generalization are elucidated by working out details for the clinical trial problem. Optimality equations have also been derived, and we focus on solving a particular simple example. There are two treatments, and the patients' lifetime after the treatments are assumed to be exponentially distributed. One of the treatments has a known expected lifetime λ. The other has an unknown expected lifetime θ^{-1}, where θ has a Gamma (η, τ) prior distribution. The random times between any two continuous treatments of the patients follow the exponential distribution with known mean σ^{-1}. We then look for the optimal strategies and detailed solutions. These are described by the function $\triangle$, which is the advantage of the unknown treatment over the known treatment. We show how this function determines the optimal strategies by specifying some break-even values of the parameters. The monotonicity property of the optimal strategies is discovered, and

the solutions for the case of two allocations of treatments is worked out and discussed in detail. Finally we corrected some errors in Wang's[40] simulations and use R to rewrite the code for the simulations. We compare our results with Wang's, and verify that in the general case the myopic strategy is not optimal at the initial phase but gradually dominates the proposed strategy. We also verified the changing trends in the function $\triangle$ and found a good lower bound $b^* = 0.93$ for searching the optimal strategy. We observe the changes that occur after altering our parameters. In particular, when we increase the total number of patients we find that the myopic strategy completely dominates the proposed strategy, and does so earlier. A larger number N can provide more initial history information for the unknown treatment. The increase in the range of advantage values at the initial phase indicates that the proposed strategy has a huge advantage. However, more adequate information has also led the myopic strategy to quickly improve its accuracy. In addition, with more selected patients there is a higher risk. If the unknown treatment is no better than the known treatment in terms of immediate payoff and the difference is bounded by a calculated risk, then it is still optimal to select the unknown treatment in order to benefit from information gathering. How to control this calculated risk is an avenue of further research. By observing the result of reducing the value of σ, we find that longer censoring times lead to more information being collected in the unknown treatment. More valid information improves the efficiency of both strategies, but the myopic strategy always improves at a faster rate and keeps the most efficient solution. Hence, there is no major difference at the initial phase if we use a smaller number of σ, but it should cause the velocity of the advantage values to decrease faster. On the other hand, the velocity of the advantage values decrease almost at a constant rate when we reduce λ. Although the myopic strategy dominates the proposed strategy earlier when either σ or λ is reduced, the biggest difference is that reducing λ causes the advantage values to decrease at the initial phase. A smaller λ should lead more strategies to choose the unknown treatment at the initial phase, and the gap must decrease when more patients receive the same treatment.

In our simulation, we used a uniform sequence for discount factors to make the calculation quick and easy. In practice, the uniform sequence is just a special case of geometric sequence. Since the geometric sequence more universal in real life, so a general geometric sequence will apply to our simulation in the future research.

Bibliography

[1] Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association* 58:365-384.

[2] Athreya, K. B., Karlin, S. (1968). Embedding urn schemes into continuous time branching processes and related limit theorems. *Ann. Math. Statist.* 39, 1801-1817.

[3] Berry, D. A., Fristedt, B. (1985). *Bandit Problems - Sequential Allocation of Experiments.* London: Chapman and Hall.

[4] Berry, D. A., Eick, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statist. Med.* 14:231-246.

[5] Bai, Z. D., Hu, F., Rosenberger, W. F. (2002). Asymptotic properties of adaptive designs for clinical trials with delayed response. *Annals of Statistics* 30:122-139.

[6] Colton, T. (1963). A model for selecting one of two medical treatments. *Journal of the American Statistical Association* 58:388-400.

[7] Cheng, Y., Berry, D. A. (2007). Optimal adaptive randomized designs for clinical trials. *Biometrika* 94:673-689.

[8] Berry, D. A., Stangl, D. (1996). *Bayesian Biostatistics.* Statistics: A Series of Textbooks and Monographs. *CRC Press.*

[9] DeMets, D. L. (2012). Current development in clinical trials: issues old and new. *Stat. Med. 31.* 2944-2954.

[10] D'Agostino, R. S., DeMets, D., Friedewald, W., Goodman, S., Witte, J., Geller, N. L. (2012). The future of clinical trials: a panel discussion. *Stat. Med. 31* 3068-3072.

[11] Eick, S. G. (1988). The two-armed bandit with delayed responses. *Ann. Stat* Vol.16, No.1, 254-264.

[12] Michael, M., John, R., Malcolm, C. (2004). Fortran 95/2003 Explained. *Oxford.*

[13] Gittins, J. C. (1989). Multi-armed Bandit Allocation Indices. *John Wiley and Sons, Chichester.*

[14] Hu, F., Zhang, L.-X. (2004). Asymptotic normality of adaptive designs with delayed response. *Bernoulli* 10:447-463.

[15] Hu, F., Rosenberger, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. Hoboken, NJ: John Wiley and Sons.

[16] Hu, F., Rosenberger, W. F., Zhang, L. X. (2006). Asymptotically best response- adaptive randomization procedures. *Journal of Statistical Planning and Inference* 136:1911-1922.

[17] Hu, F., Zhang, L. X., Cheung, S. H., Chan, W. S. (2008). Doubly-adaptive biased coin designs with delayed responses. *Canadian Journal of Statistics* 36:541-559.

[18] Hinkelmann, K. (2012). *Design and Analysis of Experiments*. New York: Wiley.

[19] Ivanova, A. V. (2003). A play-the-winner type urn model with reduced variability. *Metrika* 58:1-14.

[20] Lee,E. T. Wang, W. Y. (2003). Statistical Methods for Survival Data Analysis. *John Wiley and Sons, inc.*

[21] Lee, J. J., Chu, C. T. (2012). Bayesian clinical trials in action. *Stat. Med. 31*. 2955-2972.

[22] Oksendal, B. (2003). Stochastic Differential Equations: An Introduction with Applications. *Springer, Berlin.*

[23] Pullman, D., Wang, X. (2001). Adaptive designs, informed consent, and the ethics of research. *Control clinical Trials* 22:203-210.

[24] Piantadosi, S. (2005). *Clinical trials. A methodologic perspective.* Wiley-Interscience.

[25] Prinja, S., Gupta, N., Verma, R. (2010). Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine* 35(2): 217-221.

[26] Pong, A., Chow, S. C., eds. (2011). *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. Boca Raton, FL: CRC Press.

[27] Pappas, V., Adamidis, K., Loukas, S. (2012). A Family of Lifetime Distributions. *International Journal of Quality, Statistics, and Reliability.*

[28] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58:527-535.

[29] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* 12 1151-1172.

[30] Rosenberger, W. F., Lachin, J. M. (2002). Randomization in Clinical Trials. *Theory and Practice*. New York: John Wiley and Sons.

[31] Rosenberger, W. F. (2002). Randomized urn models and sequential design. *Sequential Analysis* 21:1-41.

[32] Rosenberger, W. F., Sverdlov, O., Hu, F. F. (2012). Adaptive randomization for clinical trials. *J. Biopharm. Statist* 22: 719-736.

[33] http://www.r-project.org/.

[34] Spiegelhalter, D. J., Jones, D., Abrams, K. (2000). Bayesian methods in health technology assessment: a review. *Health Technology Assessment* 4(38), 1-130.

[35] Sun, R., Cheung, S. H., Zhang, L. X. (2007). A generalized drop-the-loser rule for multi-treatment clinical trials. *Journal of Statistical Planning and Inference* 137:2011-2023.

[36] Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Appl. Stoch. Models Bus. Ind.* 26 , no. 6, 639-658.

[37] Scott M. B., Bradley P. C., Lee, J. J., Peter M. (2011). Bayesian Adaptive Methods for Clinical Trials. *Chapman and Hall.*

[38] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in the view of the evidence of the two samples. *Biometrika* 25:275-294.

[39] Wei, L. J., Durham, S. D. (1978). Randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association* 73:838-843.

[40] Wang,X. (1995). *Sequential Selections of Treatments with Delayed Responses.*

[41] West, M., Harrison, J. (1997). Bayesian Forecasting and Dynamic Models. *Springer, New York.*

[42] Wang, X. (2002). Asymptotic properties of bandit processes with geometric responses. *Statist. Prob. Lett.* 60:211-217.

[43] Wang, X., Tan, Q., Bicks, M. G. (2011). Clinical trials with exponential survival times. *Communications in Statistics.*

[44] Wang, X., Bickis, G. M. (2003). One-armed bandit models with continuous and delayed responses. Math. Methods Oper. Res. 58, no. 2, 209-219.

[45] Yi, Y., Wang, X. (2008). *Asymptotically efficient estimation in response adaptive trials. J. Statist. Plann. Infer.* 138:2899-2905.

[46] Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 64:131-146.

[47] Zhang, L.-X., Chan, W. S., Cheung. S. H., Hu, F. (2007). A generalized urn model for clinical trials with delayed responses. *Statistica Sinica* 17:387-409.

[48] Zhang, L.-X., Hu, F., Cheung. S. H., Chan, W. S. (2011). Immigrated urn models theoretical properties and applications. *Annals of Statistics* 39:643-671.

Appendix A

Appendix

```
## Nsimulation is a program designed by R. The objective is to see the difference of the total
##  expected lifetimes between proposed strategy and myopic strategy. This program considers the
##  motivated example with any number of selected patients, any number of sigma, any number of
##  known expect lifetime and any number of initial lower bound to make the proposed strategy
##  consistently better than the myopic strategy.
Nsimulation<-function(npatient,sigma,lambda,cstar){
 #initial simulation replication 10000 times
 ntrial<- 10000
 #initial total number of censored patients
 Nbank<-npatient
 #initial total number of deaths
 Ntau<-npatient-1
 #initial expect value of known treatment lambda
 Lambda<-rep(ntrial,(1/lambda))
 #initial proposed matrix size
 proposed<-matrix(0.0,Nbank,Ntau)
 #initial myopic matrix size
 myopic<-matrix(0.0,Nbank,Ntau)
 #initial delta matrix size
 delta<-matrix(0.0,Nbank,Ntau)
 #initial expect matrix size
 expect<-matrix(0.0,Nbank,Ntau)
 for(m in 1:Nbank){
  for(n in 1:Ntau){
   #initial deaths value for proposed strategy
   tau<-rep(n+1,ntrial)
   #initial deaths value for myopic strategy
   taustar<-tau
   #initial censored patients value for proposed strategy
   nbank<-rep(m,ntrial)
   #initial censored patients value for myopic strategy
   nbankstar<-nbank
   #initial history
   his<-((nbank+tau)>npatient)+(1-((nbank+tau)>npatient))*((l==npatient)
                              +(1-(l==npatient))*(cstar+(1.0-cstar)*(nbank+tau)/npatient))
   #initial eta for proposed strategy
   eta<-his*lambda*(tau-1)
   #initial eta for myopic strategy
   etastar<-eta
   #cumulate total time of two strategies from the first selected patient
   for(l in 1:npatient){
     #use rgamma to generate theta which is unknown hazard rate
     theta<-rgamma(ntrial,tau,eta)
     #generate parameter of lifetime distribution for unknown treatment
     param<-lapply(as.list(1:ntrial),function(a){c(theta[a],(nbank+1)[a])})
     #generate lifetimes for (nbank+1) censored patients
     life<-lapply(param,function(a){rexp(a[2],a[1])})
     #generate parameter of lifetime distribution for known treatment
     paramstar<-lapply(as.list(1:ntrial),function(a){c(Lambda[a],(nbankstar+1)[a])})
     #generate lifetimes for (nbankstar+1) censored patients
     lifestar<-lapply(paramstar,function(a){rexp(a[2],a[1])})
     #generate the censoring time
```

```
        censor<-rexp(ntrial,sigma)
      #generate survival lifetime of proposed treatment
        survival<-mapply(function(a,b){sapply(a,function(c){min(c,b)})},life,censor)
      #update eta for proposed treatment
        eta<-eta+colSums(survival)
      #generate survival lifetime of myopic treatment
        survivalstar<-mapply(function(a,b){sapply(a,function(c){min(c,b)})},lifestar,censor)
      #update eat for myopic strategy
        etastar<-etastar+colSums(survivalstar)
      #generate deaths number of proposed strategy
        death<-mapply(function(a,b){a<=b},life,censor)
      #update deaths number and censored patients number for proposed strategy
        tau<-tau+colSums(death)
        nbank<-nbank+1-colSums(death)
      #generate deaths number of myopic strategy
        deathstar<-mapply(function(a,b){a<=b},lifestar,censor)
      #update deaths number and censored patients number for myopic strategy
        taustar<-taustar+colSums(deathstar)
        nbankstar<-nbankstar-colSums(deathstar)
      #update information of history for next stage
        his<-((nbank+tau)>npatient)+(1-((nbank+tau)>npatient))*((l==npatient)
                                  +(1-(l==npatient))*(cstar+(1.0-cstar)*(nbank+tau)/npatient))
      #update criterion of eat for proposed strategy
        crieta<-his*lambda*(tau-1.0)
      #update criterion of eat for myopic strategy
        crietastar<-1*lambda*(taustar-1.0)
      #update total survival lifetime of proposed strategy
        unknown<-unknown+lambda*(eta<crieta)+(eta/(tau-1.0))*(eta>=crieta)
      #update total survival lifetime of myopic strategy
        known<-known+lambda*(etastar<crietastar)+(etastar/(taustar-1.0))*(etastar>=crietastar)
        }
      proposed[m,n]<-mean(unknown)
      myopic[m,n]<-mean(known)
      delta[m,n]<-proposed[m,n]-myopic[m,n]
      expect[m,n]<-max(proposed[m,n],myopic[m,n])
      }
    }
}
```