Upgrading Report

(from MPhil to PhD)

Proposed title: Designing a Series of Clinical Trials

by

Siew Wan Hee

University of Warwick, Warwick Medical School

October 2009

Contents

List of Tables

List of Figures

Declaration

I am aware of University regulations governing plagiarism and I declare that this upgrading document is all my own work except where I have stated otherwise.

Signed: ________________________________

Date: ________________________________

Abstract

In the development of a new drug, the design of clinical trials is usually based on consideration of a single trial at a time. Essentially, resources such as patients and money are finite and limited. As such, the decision from each trial will affect subsequent trials either in the design of the next trial or allocation of resources to other trials. The aim of this repot is to regard a clinical trial as part of a series of trials. The design of the clinical trial will incorporate how to allocate resources, i.e. patient and cost, into each trial such that the use of these resources is optimised while at the same time the power of the whole series of trial is maximised. The proposed approach uses a combination of frequentist, i.e. hypothesis inference, and Bayesian methods to find the smallest sample size to maximise the expected power of each trial and minimise the cost of a series of trial. The structure of this report is as follows. Chapter 1 introduces the types of clinical trials, Chapter 2 introduces the statistical terms and notation frequently encountered in the sample size determination literature and used in this report. Chapter 3 presents a review of literature on sample size determination for phase II clinical trials. Chapter 4 presents the designs of a series of clinical trials, Chapter 5 proposes further work and extensions to be done and finally, Chapter 6 details the courses that I have attended and my plans for the next two academic years.

Chapter 1

Clinical trial

A *clinical trial* is an experiment conducted in human beings to study and assess the effect of an intervention. The intervention could be a new drug, medical device, procedures, etc. In this report, only clinical trials involving a drug or a combination of drugs which is also referred as a *regimen* will be discussed. As such terms such as "therapy", "treatment" and "drug" will be used interchangeably throughout this report.

1.1 Characteristics of clinical trial

A clinical trial is performed prospectively and not retrospectively [Friedman et al., 1998]. Patients in the trial are followed and observed directly forward from the time of initiation of the trial. The primary objective of a clinical trial is to *compare* the effect of an experimental treatment with a control treatment. A control treatment could be a non-active intervention which means that the "treatment" is a placebo or no intervention at all or due to ethical reason the control treatment could be an “existing established effective treatment” [Fitzpatrick, 2005].

Patients are recruited and assigned to either the experimental or the control group. At baseline, patients recruited to either group should be sufficiently similar. However, in nature, patients are not uniform physiologically and thus, *randomisation* ensures that unknown factors that could influence the effects of treatment are reasonably well balanced between the treatment arms. Therefore, to ensure that the comparison can be done confidently patients are randomised to the treatment arms.

Biases can be introduced either knowingly or unknowingly in clinical trials. For example, investigators may selectively choose patients with better prognosis to the experimental arm or patients may have “convinced” themselves that they are faring better if they have known that they were given the experimental treatment. Henceforth, to minimise these biases and to ensure an objective assessment of the effects of treatments, *blinding* is introduced. A single blind trial is one in which the patient does not know what treatment has been assigned to him/her. A double blind trial is one in which neither the patient nor the investigator or the assessor know what treatment has been assigned. Some trials are triple blind. In such trials, the statistical analyst is also blinded to the treatment that has been assigned to the patient.

The trial will only be “unblinded” after the final analyses and conclusions have been made. In the words of Lilienfeld [1982], these measures are introduced to achieve the goal of *ceteris paribus*, that is, “all other factors being equal”.

1.2 Brief history of clinical trial

The precise definition and characteristics of a clinical trial we have today are formalised rather recently compared to the long history of medical research. Lilienfeld [1982] chronicled the development and evolution of the idea of clinical trial we know today. He quoted the earliest recorded account of a comparative study in the first chapter of the Book of Daniel from the Old Testament. From verses 12 to 15 Daniel proposed that two groups of servants to be given different diet for ten days. One group would have the same rich diet as the king's and another with only vegetables and water to drink. At the end of ten days, the latter group appeared fairer and healthier than the former.

A more modern written record of comparison of treatments was the famous experiment conducted by James Lind to examine the treatment of scurvy for sailors onboard. On a sea voyage on board *Salisbury*, twelve sailors with scurvy were divided into six groups with two sailors in each group. Six treatments were evaluated and they were: (1) cider, (2) diluted sulphuric acid, (3) vinegar, (4) seawater, (5) a mixture of several foods including nutmeg and garlic, and (6) oranges and lemons [Hackshaw, 2009]. After six days one of the sailors given oranges and lemons was fit for duty whereas the other one showed the most signs of improvements compared to the other ten patients. Although Lind noted the value of the fruits, they were not recommended as treatment for the disease mainly because of the expense of such fruits compare to "pure dry air" which was the recommended treatment on board after the trial.

Lind's experiment had an important feature of clinical trial: comparing two or more interventions. However, he admitted that there was "favouritism" in assigning the patients to the treatment arms. "Two of the worst patients ... were put under a course of sea-water." Lind might have had a belief that the sea water might be the best treatment. In today modern time, randomisation would be able to eliminate the bias in selecting patients to a treatment arm that is perceived to be more effective.

Randomisation was first proposed by R. A. Fisher in the experimental study of agriculture. Plots of crops were randomised to receive different treatments and he argued that the randomisation would simulate independence and as such, the statistical analysis which is often based on the assumption of independence would then be valid.

One of the earliest clinical trials that adopted a randomisation procedure was a trial conducted in the University of Minnesota where at the beginning of the university term, students were asked to volunteer to participate in a cold vaccination trial. They were then randomly assigned to either the experimental or the control group. Although the students and the attending physicians were blinded of the assigned treatment, the randomisation assignment was done systematically where if a student was assigned to an experimental arm then the following student would be assigned to the control group. A problem with such systematic randomisation is that it is easy to predict the treatment the next student would receive.

One of the first reported clinical trial that used random numbers to randomise patients into treatment groups is a trial involving treatment of streptomycin in pulmonary tuberculosis by the Medical Research Council of Great

Britain in 1948. This trial was a single blind trial, that is, only the two radiologists assigned to read the x-rays independently were blinded of the treatments assigned to the patients. Both patients and the attending physicians were told of the treatment that they were receiving.

Following the success of the streptomycin trial, more and more comparative trials started to incorporate randomisation and blinding procedures. Indeed the characteristics of clinical trials mentioned by Lilienfeld [1982]; comparative, randomise and blinded have become the *gold standard* for clinical trials. For further readings, please refer to Machin et al. [2006, Chap. 1] and Meinert and Tonascia [1986, Chap. 1] for summary of the history and development of clinical trials, and Bull [1959] for a scholarly account on the historical development of clinical trials.

1.3 Designs of clinical trial

The clinical development of a new drug is usually divided into four phases [Machin et al., 2006, pp. 13-37]. The nomenclature of phase I, II, III and IV has thus been developed for the purpose of classifying the objective and goal of each phase of testing in the drug development programme. Broadly, the purpose of a phase I trial is to study the human pharmacology and safety, phase II is to explore therapeutic activity, phase III is to demonstrate or confirm the therapeutic activity observed in phase II, and finally phase IV is conducted after drug approval to understand better the usage of the therapeutic agent according to the approved indication [Fitzpatrick, 2005].

Phase I and II trials are collectively known as early phase trials and they are typically conducted to ascertain three objectives: (1) to document the experience (such as the dosage, the route of administration and schedule of the drug, and the type of patient) and develop a protocol for reproducibility, (2) to learn any unacceptable toxicity and adverse event, and (3) to determine if the treatment is efficacious before it is brought forward for more vigorous testing in larger phase III trials [Schoenfeld, 1980].

Traditionally and in an ideal situation the development of a new drug would go through a series of clinical trials sequentially through phase I to IV. This is because the results from the preceding phase are used to motivate the design of the next phase. In practice, the development plan may not go through the same sequence. It is rather common for the results from a phase II exploratory study to prompt additional human pharmacology studies or to modify the strategy of drug administration or to lead to more studies to investigate the dose-response relationship. Or the results from a phase III trial may prompt another phase III trial by narrowing the disease population.

1.3.1 Phase I

Friedman et al. [1998] defined a phase I study as an introductory study where a drug is first tested in humans. In most diseases, healthy volunteers are recruited to participate in the phase I trial. Hence, the term “participants” will be used as the subjects of phase I trial. An exception is oncology trials where the treatments are highly toxic, and only cancer patients are recruited for phase I trial. Usually, the patients have already tried and failed on the existing standard therapies. Green et al. [2003] gives a general overview on the design of oncology trials and Crowley and Ankerst [2006] provides more technical details.

The main objective of a phase I trial is to estimate the maximum dose level that is acceptable for a participant without causing unacceptable toxicity. This dose is conventionally known as the *maximally tolerated dose* (MTD) and the unacceptable toxicity is known as the *dose limiting toxicity* (DLT). A commonly used phase I design is known as the "3 + 3" design. In this design, a few doses are identified for consideration. Let d_1 be the starting dose which is extrapolated from animal studies and the highest dose level planned for the trial be d_m. Supposed that the other dose levels in between are $d_2, d_3, \ldots, d_{m-1}$. A cohort of three participants will be recruited to a dose level, d_i $(i = 1, 2, \ldots, m)$, and if no patient experiences any DLT then another cohort of three participants will be recruited to dose level, d_{i+1}. If however, one participant experiences any of the predefined DLT, then another cohort of three participants will be recruited to the same dose level, and if no further DLT is observed among this new cohort of participants the dose level is escalated. If at any dose level at least two participants experience any DLT, then the trial will stop and the dose level preceding it will be declared MTD (Figure 1.1). If the dose level d_m is reached with no DLT observed then the investigator may declare d_m as the MTD or another phase I trial may be initiated.

Note that the primary objective of a phase I trial is not comparative in nature and hence, there is no statistical test involvement. Due to the uncertainty from the escalation and de-escalation of dose level, the total number of patients needed is hard to fix during the planning and designing

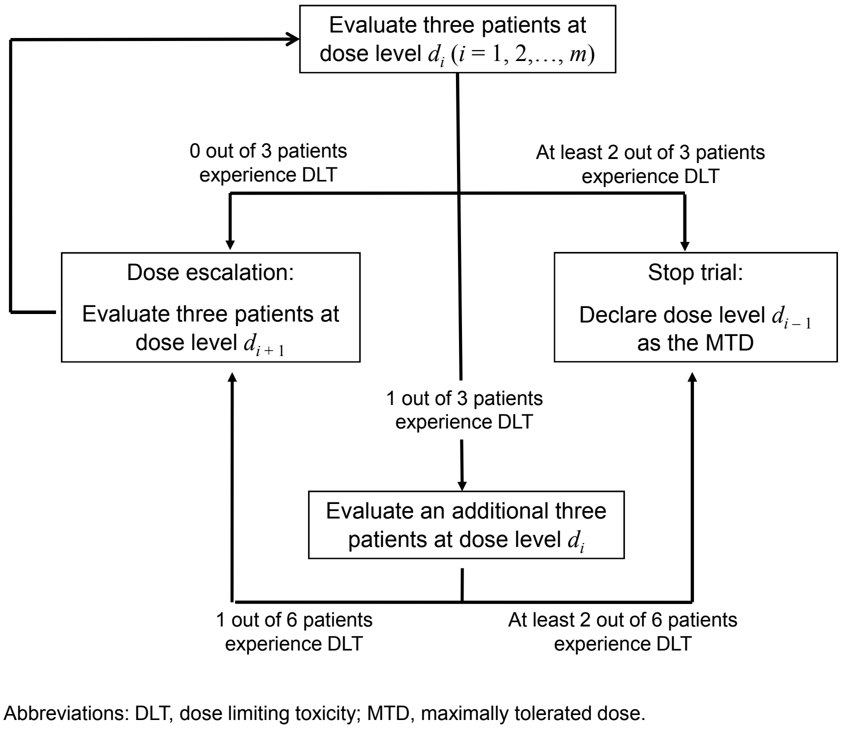

Evaluate three patients at dose level d_i ($i = 1, 2, \ldots, m$)
0 out of 3 patients experience DLT
At least 2 out of 3 patients experience DLT
Dose escalation:
Evaluate three patients at dose level d_{i+1}
Stop trial:
Declare dose level d_{i-1} as the MTD
1 out of 3 patients experience DLT
Evaluate an additional three patients at dose level d_i
1 out of 6 patients experience DLT
At least 2 out of 6 patients experience DLT
Abbreviations: DLT, dose limiting toxicity; MTD, maximally tolerated dose.

Figure 1.1: A "3 + 3" design to establish a maximum tolerated dose (MTD) in a phase I clinical trial.

stages of the clinical trial. However, the maximum number of patients is often stated in advance in the trial protocol. In practice, usually three to eigth dose levels are investigated in a trial. Hence, the maximum number of patients ranges from 18 (a maximum of 6 patients in each dose level) to 42.

Although the "3+3" design is commonly practised in most phase I clinical trials because of its simple implementation, there are other strategies that have been recommended by other authors especially for oncology trials. As mentioned earlier, patients are recruited for phase I oncology trial instead of healthy volunteers. Sometimes the first few dose levels may be less efficacious, and the "3 + 3" design would inadvertently recruited too many patients to suboptimal doses. O'Quigley et al. [1990] presented a Bayesian design, the *continual reassessment method* (CRM), that claims to subject less patients to the low inferior dosages.

Further technical details of phase I design are available in Machin et al. [2009] and other references cited therein.

1.3.2 Phase II

Once the MTD is established, the treatment is further evaluated in a phase II trial to investigate efficacy. Both Gehan [1961] and Schoenfeld [1980] describe that only one group of patient is selected for the trial and they are usually a homogeneous group in terms of disease and stage of disease.

Although the purpose of the phase II study is to explore for efficacy, the primary endpoint may be a surrogate endpoint. For example, in oncology trials, typically the survival rate or remission rate is of importance. However, it takes a long time to capture adequately these variables and it is not feasible to have long follow up duration in a phase II trial [Schoenfeld, 1980]. Hence, a surrogate endpoint such as tumour shrinkage is used as the primary endpoint. It is generally acceptable that if the tumour shrinks by a considerable size, the patient will have longer survival. Thus, on the one hand, the tumour shrinkage may be considered as a continuous response where the change of the sum of the longest diameter is used as an endpoint.

On the other hand the tumour shrinkage may be considered as a binary response where either the sum of the longest diameter of the target lesion decreases at least 30% according to the Response Evaluation Criteria in Solid Tumors (RECIST) criteria [Therasse et al., 2000] and hence considered as a *success* or the dimension remains unchanged or increases by at least 20% then it is considered as a *failure*.

The phase II trial is the first trial in a drug development programme to compare the efficacy of the drug formally. As such, the problem is formulated statistically into a hypothesis test. However, it is not necessary to have a control arm in a phase II trial. In an example given by Gehan, suppose that there is a difference of 10% between standard and experimental treatments when the standard treatment shows 5% effectiveness. Thus, by controlling the false positive (claiming the new treatment is effective when in fact it is not which is also known as type I error) and false negative (claiming the new treatment is not effective when in fact it is which is also known as type II error) rates at 0.05 and 0.10, respectively, the sample size required is 191 for each arm. To have so many patients in a phase II trial is a waste of resources, namely, patients, effort, time and money. Suppose that only the experimental treatment is considered in a phase II trial then the total number of patients will be reduced from 382 to 79 [Schoenfeld, 1980].

In addition, the outcome of a phase II trial is to choose between to bring the treatment forward for further testing in confirmative phase III trial and to stop the treatment from further testing. Therefore, it is important to stop futile trials as quick as possible and ensure that effective drugs are given the go-ahead for further testing with some degree of confidence. This decision is rather straightforward and it is another reason why a standard arm is not necessary for statistical analysis such as those found in *proof-of-concept* phase III trial. Therefore, the hypothesis test compares the new treatment to a historical control.

Some of the common designs of phase II trials are *single-stage* and *two-stage designs*. These are explained in more detail below. Other authors have proposed *multiple stage designs* [Chen, 1997, Ensign et al., 1994] and these are usually extension of the single- and two-stage designs. Most of the works on the designs of phase II trials originate from evaluation of anticancer therapies. Also, most of the approaches consider a binary response as mentioned earlier. Briefly, in the single-stage design, a sample size is determined for the trial and analysis is done after all the data have been gathered. A two-stage design is akin to a single-stage that is split into two. In the first part of the two-stage which is simply known as first stage, a total of n_1 patients is recruited. Then the data are analysed and if a minimum number of successes is observed then a further n_2 patients are recruited to pinpoint the accuracy of the effectiveness of the drug. If however, the minimum number of successes is not observed, then the drug will be rejected and not recommended for further testing.

Another type of design in phase II is a *sequential design* where the hypothesis is repeatedly tested [Mariani and Marubini, 1996]. There are two approaches in this design. In the first approach, each patient is recruited sequentially, and an analysis is performed when the outcome from each new patient is available and added on to the accummulated data. There are three possible decisions from each analysis: (1) stop the phase II trial and declare the new treatment is not worthy for recommendation to phase III trial, (2) stop the phase II trial and recommend the new treatment to phase III trial, and (3) continue with the phase II trial and recruit the next patient.

In another approach, a group of patients are recruited and their outcomes are pooled for analysis. From the result, the same choices of decisions are available, (1) abandon the phase II trial, (2) proceed to phase III trial, and (3) recruit another group of patients and use the accumulated data for the next analysis to test the hypothesis.

Although a phase II trial does not necessitate a control arm, there are some designs especially in oncology trials where there are two or more treatment arms in a trial. The treatment arms involved in a randomised phase II trial are all experimental treatments. The aim of randomised phase II trial is not to compare definitively but to identify a promising treatment for further testing. Such designs are known as *selection designs* [Simon et al., 1985].

A review by Mariani and Marubini [1996] provides a general overview of phase II clinical trials especially in the application of oncology trials. Other texts by Machin et al. [2009] and Stallard [2008] give very good examples of designs and statistical methods of phase II clinical trials.

The sample size determination of phase II trial can be broadly categorised into two methods: (a) frequentist, and (b) Bayesian [Adcock, 1997]. The former is based on an inferential method where a statistical method is used to infer the hypothesized efficacy of the treatment from the observed patients responses. The latter method can be further classified into two groups of techniques, namely inference on the treatment efficacy and a decision problem where the optimal course of action based on the merits and demerits of each viable decision. The sample size determination of phase II trial will be discussed in further details in Chapter 3.

1.3.3 Phase III

A phase III trial is a definitive clinical trial that is comparative, randomised and blinded. It is a large confirmatory trial where the results are submitted to regulatory authorities for drug approval and is conducted with at least a control and intervention groups. In large trials such as phase III trial, the probabilities of both type I and II errors are minimised because the sample size is large.

Due to the large sample size required in a phase III trial, it is mainly conducted concurrently by a few centres and sometimes known as multicentre trials. The advantage is the possibility of wider patient population recruitment and a broad range of clinical settings that is more typical of future use. There may arise some situations in phase III trials where it is imperative to check the assumption of the original design, the efficacy and/or safety of the experimental treatment before the formal completion of the clinical trial. Hence interim analysis (analyses) is (are) built into the trial protocol with stopping rules defined to allow the trials to stop early if the superiority or inferiority of the experimental treatment is established.[ICH E9, 1998]

The sample size determination of phase III trial is traditionally based on the classical approach of hypothesis testing. The Bayesian sample size determination is less popular in phase III trial mainly because it is usually associated with sequential procedures which may be unable to give a fixed sample size at the design stage of the clinical trial, and some of its computations require complicated algorithm. [Pezeshk, 2003] However, in the recent years the developments of researches based on Bayesian sample size determination has increased and gained further stronghold in clinical trials. Adcock [1997] gave a very good review of sample size determination for both classical frequentist and Bayesian approaches. Although the review is meant for general applicability, it is easily adapted for clinical trial settings. Pezeshk [2003] on the other hand, reviewed Bayesian methodology used in clinical trials. Both authors have claimed that their reviews are not meant to be comprehensive and strongly encouraged readers to refer to references cited therein for further thorough readings.

1.3.4 Phase IV

Phase IV trials are usually undertaken after the registration or during the registration of a drug to monitor and discover more about the safety of the drug. Sometimes, the trial also assess for efficacy in different populations. The sample size in such a trial is very large and may not have a control arm.

In the following chapters, we will not discuss the designs of phase I and IV trials. Hence, in particular we shall assume that the dosage and safety issues of the new drug or treatment regimen has been addressed in phase I trials.

Chapter 2

Statistical terms and notation

One of the key issues in the planning of a clinical trial is the sample size determination. Following the guidelines in 'Statistical Principles for Clinical Trials' of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH E9), "the number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed." [1998] In the determination of an appropriate sample size, it is not a simple matter of picking a number from a sample size table but it necessitates various information such as the objective of the trial, patient population, the maximum allowable error rates, accrual rate of patient, etc. In this chapter, we will introduce and explain the statistical principles and methodology commonly occurred in the sample size determination of clinical trials.

2.1 Notation

The *sample space* is a mathematical set that represents all possible observations in a situation under a specific trial. The sample space can be either *continuous* or *discrete.* If the sample space is continuous, it is said to have some measure such as length and if it is discrete then it is finite or countably infinite.

In each trial an outcome of interest is made and sometimes instead of being interested in the outcome itself, we are interested in some function of the outcome. This function is known as a *random variable* which is defined on the sample space. Suppose X is a random variable, then each possible *observation* in X is denoted as x. In this report, X is assumed to take values on the real line.

There exists an inherent variability that is beyond the observer's control when observing an outcome. This variability most often can be described by a *probability distribution.* Assume that for a discrete sample space, $x = (x_1, x_2, \ldots)$ is an ordered set of possible observations, then there is a function p defined as

$$p(x_i) = \Pr(X = x_i), \quad i = 1, 2, \ldots$$

The function $p(x_i)$ is a non-negative function ($p(x_i) \geq 0$) and the sum of all the masses associated with each element in the space is unity ($\sum_i p(x_i) = 1$). This function is called *probability mass function.* The equivalent of probability mass function for a continuous sample space is known as *probability density function* which is also a non-negative function of the real variable x such that $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$. For an interval of $[a, b]$, the probability that X

falls into the interval is the area under the density function between a and b;

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x.$$

Note that for X is continuous, $\Pr(a \leq X \leq b) = \Pr(a \leq X < b) = \Pr(a < X \leq b) = \Pr(a < X < b)$. Suppose that $b \to a$, then $\Pr(a \leq X \leq a) = \int_a^a f(x)\,\mathrm{d}x = 0$ which means that the probability that a continuous X takes on a fixed value is 0.

The *cumulative distribution function*, F, is also frequently encountered and it is expressed in terms of $p(x)$,

$$F(x) = \Pr(X \leq x) = \begin{cases} \sum_i^x \Pr(X = i), & \text{discrete distribution} \\ \int_{-\infty}^x f(u)\,\mathrm{d}u, & \text{continuous distribution} \end{cases}$$

Suppose that in a trial the probability of success is θ, we do not know what value it is except for some value between 0 and 1. This θ is known as a *parameter* and the mathematical set that θ belongs to is usually denoted by Θ. In this report, we will assume that the probability distribution is of some known form depending on some unknown parameters.

The density function is now rewritten as $p_\theta(x)$ or $p(x, \theta)$ and interpreted as the probability density of a real x when θ is the true parameter or for a discrete x, the probability of point x when θ is the true parameter. The function $p(\cdot, \theta)$ is the density function on the sample space X, and the function $p(x, \cdot)$ is the function of the parameter space Θ which is also known as the *likelihood function*. The likelihood function is an important function as it summarises all the information the observed data can provide about the parameter θ.

2.2 Discrete distribution

A random variable that only takes two values: *success* and *failure* with probabilities p and $1-p$ in a trial, respectively is known as a *Bernoulli* random variable. Supposed a numerical 1 is used to denote success and 0 a failure, then the probability mass function is,

$$p(x) = \begin{cases} p^x(1-p)^{1-x}, & \text{if } x = 0 \text{ or } x = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that n independent patients are recruited and the total number of successes, X, is observed. Then X is a binomial random variable with index n and parameter p. The *binomial distribution* is constructed from n independent Bernoulli trials and the sequence of the occurrence of successes is not important. Thus there are $\binom{n}{x}$ ways in which a total number of x successes may occur in n trials. The probability mass function is

$$p(x) = \Pr(X = x) = \binom{n}{x} p^x(1-p)^{n-x}.$$

The statement of X following a binomial distribution with (n, p) can be "rewritten" as $X \sim \mathrm{B}(n, p)$.

The *geometric distribution* is another distribution constructed from Bernoulli trials. The difference is that there is an infinite number of trials. A sequence of trials are conducted and the probability of a success is p. The trial will stop when the first success is observed. Let X be the total number of trials including the first successful outcome and following from the independence of the trials, the probability mass function is,

$$p(x) = \Pr(X = x) = (1 - p)^{x-1}p, \quad x = 1, 2, \ldots$$

Note that $\sum_{x=1}^{\infty}(1 - p)^{x-1}p = p\sum_{x=1}^{\infty}(1 - p)^{x-1}$ and by the expression giving the sum of a geometric series we have $\sum_{x=1}^{\infty}(1 - p)^{x-1} = \frac{1}{p}$, therefore, $\sum_{x=1}^{\infty}(1 - p)^{x-1}p = 1$.

2.3 Continuous distribution

The simplest continuous distribution is the *uniform distribution*. The random variable X is said to have a uniform distribution if its probability density function is given by

$$p(x) = \begin{cases} (b-a)^{-1}, & a < x < b \\ 0, & \text{otherwise.} \end{cases}$$

A special case of a uniform distribution is for $a = 0$ and $b = 1$ when $p(x) = 1, \quad 0 < x < 1$.

The *normal distribution* which is also known as the *Gaussian distribution* is the most important continuous distribution and "plays a central role in probability and statistics" [Rice, 1995]. The probability density function of a normal distribution is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/2\sigma^2},$$

where the parameters θ and σ are the mean and standard deviation of the normal density, respectively. The cumulative distribution function is

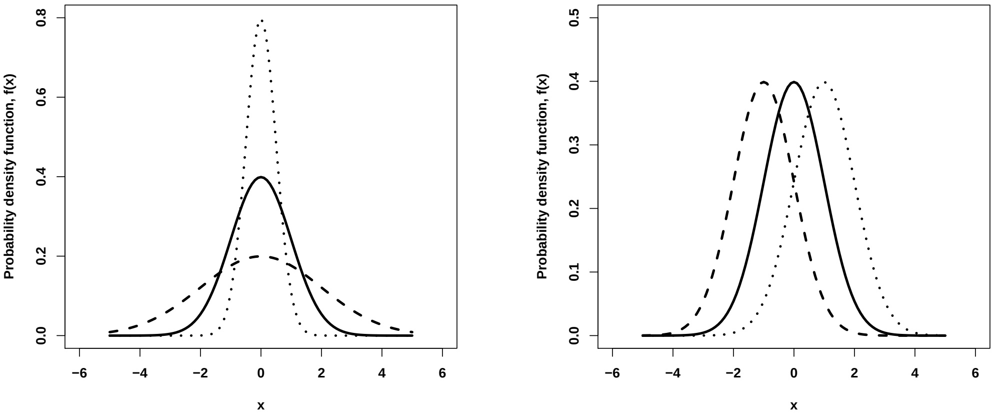
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\theta)^2/2\sigma^2}\, \mathrm{d}u$$

Suppose X is a random variable that follows the normal distribution with parameters θ and σ, the statement can be "rewritten" as $X \sim N(\theta, \sigma^2)$. The density of the normal distribution integrates to 1 in the whole space of $(-\infty, \infty)$. However, the cumulative distribution function cannot be evaluated in a closed form but has to be computed numerically.

A special case of the normal distribution is the *standard normal distribution* where the $\theta = 0$ and $\sigma^2 = 1$. Its density function is usually denoted by $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and its cumulative distribution function is denoted by $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, \mathrm{d}u$. Note that the relationship between a normal and standard normal distribution can be stated by: $p(x) = \frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)$ and $F(x) = \Phi(\frac{x-\theta}{\sigma})$.

The normal distribution when plotted in a plane of $f(x)$ against x has a bell-shaped curve (Fig. 2.1). It is symmetric about its mean, θ, and the shape of the curve, either narrow or wide, depends on the standard deviation, σ.

The *beta distribution* is a distribution that has very flexible shapes with



(a) $\theta = 0$

(b) $\sigma = 1$

Figure 2.1: Normal densities (a) σ of 0.5 (dotted), 1 (solid), and 2 (dashed), and (b)θ of -1 (dashed), 0 (solid), and 1 (dotted).

two parameters a and b. Its density function is,

$$p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1, \quad a, b > 0,$$

where $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$ is a gamma function if x is a non-integer and a simple factorial function, $\Gamma(x) = (x-1)!$ when x is an integer. As shown in Figure 2.2, the beta density has very flexible shapes, from flat to narrow curves with various values of a and b. Note that when $a = 1, b = 1$, the beta distribution is a uniform distribution.

2.4 Estimation and confidence set

Due to the inherent variability in observing an outcome in each situation, a probability distribution is used to describe the variability. However, the

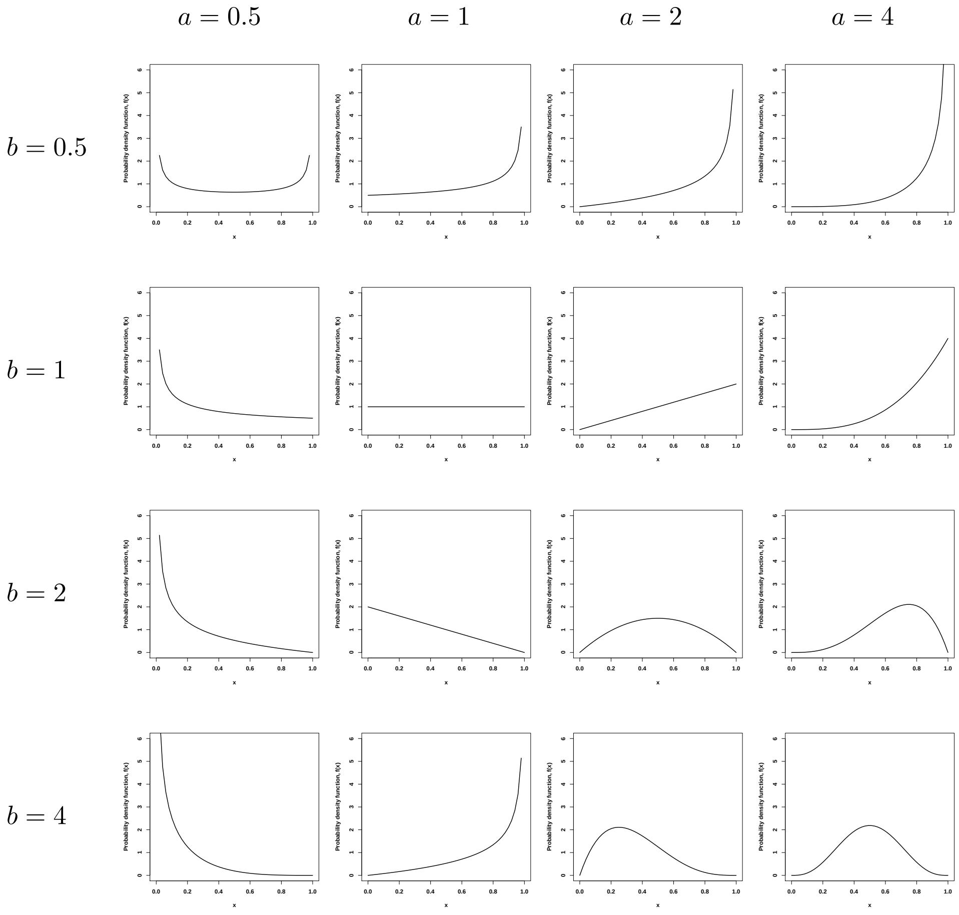


Figure 2.2: Beta densities with various values of a and b.

probability distribution is also unknown to us. The inference problem is thus to infer something of the true distribution or rather the true parameter - as we are assuming that the distribution is known with unknown parameter - from the observed outcomes. The numerical values of the observations are used to *estimate* the unknown parameter.

For example, in a trial with n independent patients and we observed either a success or failure from each of them. Suppose that the probability of success is θ, it seems reasonable that the proportion of successes, k/n, where k is the total number of successes observed is a 'good' estimate of θ. There are a few criteria that decides how 'good' an estimator is in estimating the parameter. The discussion of these criteria are rather technical and so will not be discussed here. Instead readers are encouraged to refer to more technical texts such as Silvey [1975] and Cox [2006].

Although an estimator may be a good approximate of the plausible parameter, it is unlikely that the estimate is the same as the true parameter at all times under all circumstances. Therefore, an interval of plausible parameter value is reported and this interval is also known as *confidence interval.* To construct a confidence interval, let C_L and C_U be two random variables where $C_L < C_U$. The probability that the parameter θ is within the interval (C_L, C_U) under the assumption that θ is the true parameter is

$$\Pr(C_L \leq \theta \leq C_U) = 1 - \alpha \tag{2.1}$$

The statement in (2.1) is interpreted as: the probability that the true value of θ is within this random interval is $1 - \alpha$ or equivalently for any observed x we are $100(1 - \alpha)\%$ confident that the true parameter will lie in this interval.

2.5 Statistical inference

Hypothesis testing is the theory of inferring the nature of the true parameter from the observations. A statement is necessary to imply that the true parameter θ belongs to a subset of the parameter space Θ. This statement is known as an hypothesis. The testing of the hypothesis is to use statistical methods to check if the observations are consistent with the stated hypothesis or not. Silvey [1975] summarised succintly that a statistical rule is used to assign "each possible observation to one of two exclusive categories: 'consistent with the hypothesis under consideration' and 'not consistent with this hypothesis'."

In the classical approach of hypothesis testing which is also known as the *frequentist method*, there are two hypotheses. The first is the *null hypothesis* which states that the parameter θ belongs to ω which is a subset of Θ. The other hypothesis is simply known as the *alternative hypothesis* which states that the parameter θ does not belong to the subset ω but belongs to $\Theta-\omega$. If there is only one element in ω, the hypothesis is known as a *simple* hypothesis because it is in its simplest form, and similarly, if there is only one element in $\Theta - \omega$ the alternative hypothesis is a simple alternative hypothesis. Suppose that the elements in ω and $\Theta - \omega$ are θ_0 and θ_1, respectively, the hypotheses can be formulated as

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta = \theta_A,$$

where the statement H_0 is the null hypothesis and H_1 is the alternative hypothesis. Note that the null hypothesis is always assumed to be true until proven to be otherwise.

Two possible decisions can be made based on the observed data at the end of the trial: (1) reject the null hypothesis, or (2) do not reject the null hypothesis. The decision is made on the basis of a *test statistic*, $T(X)$. The choice of $T(X)$ depends on the distribution that is assumed and the specified hypotheses. Note that a test statistic may or may not be an estimator of the parameter. If t is the observed value, the probability $\Pr(T \geq t)$ under the assumption that the null hypothesis is true is known as *p-value* and it has a known distribution under H_0 which is the uniform distribution. The *p*-value is interpreted as the probability of the statistic being as large or larger than the observed value if the null hypothesis is true.

In making any decision from the hypothesis testing, we may not always be 100% correct. Inevitably, errors may occur when rejecting or not rejecting the H_0. The *type I* error is an error incurred when the null hypothesis is rejected when it is true. Another type of error that can be incurred is the *type II* error. It is an error incurred when the null hypothesis is accepted when it is false.

The consequence of type I error is usually considered to be graver than that of making a type II error. Hence, the probability of making such error is capped by a predetermined value, generally denoted by α. Although the choice of α could be arbitrarily, it is customary to have α at small values such as 0.1, 0.05 or 0.01.

The probability of a type II error is also controlled but now it is under the assumption that the alternative hypothesis is true. The maximum allowable probability of type II error is generally denoted by β. The probability that H_0 is rejected when it is false is simply $1-\beta$. This probability is also known as the *power* of the test. Customary, β is set at some small values such as 0.2, 0.1 or 0.05, and the corresponding power is 0.8, 0.9, or 0.95, respectively.

2.6 Bayesian method

The Bayesian statistical inference technique extends the frequentist method by proposing that the parameter θ is random, that is, it has its own probability density function (probability mass function if θ is a discrete variable) which is denoted by $p(\theta)$. Note that for convenience the term *density* will be used for both continuous and discrete variables.

Before any data and evidence are available from the experiment, the investigator and statistician provide some reasonable opinion concerning the probable value of the parameter. As such, the density $p(\theta)$ is known as a *prior density*. Suppose now some data x have been observed whose probability of occurrence is assumed to depend on the random parameter θ, it is expressed as, $p(x|\theta)$. This is the likelihood function. Due to the randomness of θ the *marginal density* of x is given by the density, $p(x|\theta)$, averaged over all the possible values of θ, that is, $p_X(x) = \int_\theta p(x|\theta)p(\theta)\,\mathrm{d}\theta$.

The two sources of information (prior distribution and likelihood function) are combined to update our prior belief of the parameter θ and thus, its density is now denoted by $p(\theta|x)$. This density is known as *posterior distribution*. By Bayes theorem, the posterior distribution is estimated by

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p_X(x)}.$$

As pointed out by Lindley [1971] the terms prior and posterior are not referring to the distribution but to the relationship between the distribution of θ and the observed data x. In some circumstances, after the estimation of the posterior density we may have a final opinion of the true parameter and proceed to make a decision. In other circumstances, the new opinion may prompt further study to glean more information of the true parameter. As such, "today's posterior will become tomorrow's prior".

2.7 Bayesian decision method

Although the hypothesis testing is an inference problem, after the collection of observations, a decision is made from a choice of two. These decisions are:

d_0: The hypothesis that an unknown θ belongs to ω is true.

d_1: The hypothesis is false.

Suppose that D is the *decision space* for all the viable decisions and d is each of the decisions in D. Each decision has its consequence and "value". The "value" could be a monetary reward which is measurable in existing scale or it could be a value that has no obvious scale of measurement, such as happier feeling. However, to work on these "values", numbers are assigned and they are called *utilities.* The function of decision and parameter is called the *utility function* and is denoted by $U(d, \theta)$. A decision problem is solved by *maximising* the expected utility which is based on the posterior distribution,

$$\int_{\theta} U(d,\theta) f(\theta|x)\,\mathrm{d}\theta.$$

Instead of examining the utility, we may work on the *loss* which is the opposite of the utility and the decision problem is solved by *minimising* the loss function. For a more detailed work on the concepts and methods of decision theory, please refer to Berger [1980].

Chapter 3

Literature review

There are many therapeutic agents available for clinical trials but not all of them can be tested in large comparative trials due to limited resources such as patients, time and money. Thus, phase II trials serve to screen out nonpromising therapies. Many designs of phase II clinical trials are modelled from oncology trials. As such, the primary objective is usually to look for anti-tumour activity and a frequently used endpoint is of binary nature: success (the tumour shrinks by at least 30% according to the RECIST criteria [Therasse et al., 2000]) and failure (the tumour does not shrink or in worse case scenario it increases by at least 20%).

In this chapter, the focus of the literature review is on the design of single-arm phase II trials. Most of the designs to be discussed are based on a binary endpoint and although the motivation behind most of these designs are for anti-cancer therapies, they are easily adaptable for other diseases.

3.1 Frequentist methods

In a review by Mariani and Marubini [1996] there are three possible types of phase II designs for single-arm trials, namely, fixed-sample, sequential and multi-stage designs. The simplest design of a phase II trial is the fixed-sample design which is also known as a single-stage design where a number of patients, n, is recruited in a trial. At the end of the trial, the observed data are collected for analysis and a decision is made whether to reject the null hypothesis or not. Suppose that in a clinical trial, an observation from each patient is whether he/she is responding to the experimental drug, thus a success (S) or a failure (F). The primary endpoint is the number of successes, which follows a binomial distribution. Denoting the true probability of success by p, we wish to test the one-sided hypothesis of

$$H_0 : p \leq p_0 \quad \text{vs.} \quad H_1 : p \geq p_A$$

where p_0 is the proportion for the historical control and p_A is the minimum proportion for the new treatment to achieve to warrant further investigation.

The clinical trial is designed such that both type I and II errors (of claiming a treatment as promising when it is not and rejecting a treatment when it is effective, respectively) are capped by the maximally allowable levels,

$$\begin{aligned} \Pr(\text{reject } H_0 | H_0 \text{ is true}) &= \Pr(\text{type I error}) \leq \alpha \\ \Pr(\text{accept } H_0 | H_1 \text{ is true}) &= \Pr(\text{type II error}) \leq \beta. \end{aligned} \tag{3.1}$$

Thus, the sample size is approximately,

$$n = \left(\frac{z_{1-\alpha}\sqrt{p_0(1-p_0)} + z_{1-\beta}\sqrt{p_A(1-p_A)}}{p_A - p_0} \right)^2 \tag{3.2}$$

where $z(x)$ is the upper $100(1-x)\%$ percentile of a standard normal distribution [Schoenfeld, 1980]. Under the constraints of (3.1), there exists a value k $(0 < k < n)$ which is the minimum number of successes that need to be attained so that the investigators can decide whether to recommend the new treatment for further testing or not. The minimum number of successes k depends on n and is also known as the cut-off.

In the phase II setting, the type II error is more serious than type I error because by incurring a type II error a better drug would be denied the chance of being studied further and patients are not able to benefit from a more superior drug [Schoenfeld, 1980]. Suppose that a trial is planned to determine if the new drug can increase the response rate by 20% from the current response rate of 45% from the standard drug. By minimising the type I error rate at $\alpha = 0.25$ and the type II error rate at $\beta = 0.10$, a minimum number of 24 patients is needed to observe a difference of 20%. If more than 50% response rate was observed then the new drug should be recommended for further testing.

The sample size determination in (3.2) is the same as the one used in a usual phase III trial. To ensure that the sample size estimation is smaller for a phase II design, the difference between experimental treatment and historical control success rate is set large or the false positive and negative rates are set higher.

Note that the sample size estimation based on (3.2) is an approximation based on the normal distribution although the proportion of success has a binomial distribution. For the example above, the actual type I and II errors based on the binomial distribution will not be exactly 0.25 and 0.10, respectively. Based on the 50% cut-off rate, if the true response rate was 0.45 then the probability of observing at least 13 successes is 24% and if the true response rate was 0.65 then the probability of observing at least 13 successes is 91%. Therefore, the actual type I and II error rates are 24% and 9%, respectively. This problem is remedied by A'Hern [2001] who used the exact binomial distribution to compute the sample sizes.

A model of sample size determination according to Gehan's approach is an example of two-stage design although when it was first proposed, Gehan referred it as preliminary and follow-up trials where preliminary is what we would call the first-stage, and follow-up is the second-stage [Gehan, 1961]. An example according to Gehan's model is, suppose that the drug could be effective in 20% or more of the patients then there would be more than 95% probability of observing at least one success in 14 consecutive patients enrolled into the trial. This is assuming that the true effectiveness is 20% and both α and β are fixed at 5%. If there was no success observed among the 14 patients, the drug will be rejected and not recommended for further testing. If however, at least one success was observed, then more patients would be recruited to pinpoint the effectiveness of the drug. The additional number of patients to be recruited following the initial observed success(es) is chosen so that the true probability of success "is estimated with given precision, i.e. standard error."[Gehan, 1961]

Gehan's method requires the number of successes in the first stage to determine the sample size needed for second stage. Thus, the total number of patients needed cannot be determined at the design stage. Simon [1989] proposed two two-stage designs: optimum and minimax designs. In both designs, the number of patients needed for stage one (n_1) and two (n_2) are determined at the stage of design. Also, the cut-offs for stage one (k_1) and the whole trial (k) are determined in the design stage. The decision whether to proceed to stage two of the trial is based upon the minimum number of successes observed at the end of stage one. If the true probability of response is p and the number of successes observed is k_1 or less at the end of stage one, then the trial will end early. The probability of terminating the trial at stage one is $\Pr(X \leq k_1) = \sum_{x=0}^{k_1} \binom{n_1}{x} p^x (1-p)^{n_1 - x}$ and the expected total sample is, $\mathrm{E}(n) = n_1 + n_2 \Pr(X \leq k_1)$.

In both Simon's optimal and minimax designs, the hypothesis is a one-sided hypothesis. In his designs, the sample sizes are determined under the constraints of α and β error rates by minimizing the expected sample size, $\mathrm{E}(n)$, assuming that the true response rate is p_0. In the optimal design, the number of patients needed for stage one (n_1) is kept to a minimum to ensure that not many patients are subjected to an inferior drug. On the other hand the minimax design is to choose the smallest maximum total sample size $n(= n_1 + n_2)$ that satisfies the design error probability constraints.

3.2 Bayesian methods

In practice, although the response rate of the standard treatment or the historical control should be fixed, most often, investigators are uncertain of an exact value. Therefore, usually a range of values of p_0 is given. Due to the uncertainty in the value of p_0, Thall and Simon [1994] argued that it is then realistic to explicitly consider p_0 as random during the planning of the clinical trial and in the interpretation of the result at the end of the trial.

In addition, the decision to be made from a phase II trial is to either recommend the drug for further study or not. As such, it seems to be intuitive to adopt Bayesian method in determining the sample size where the emphasis is not to reach for a correct conclusion under the constraint of low error rates but to reach a best course of action (see Brunier and Whitehead [1994] and Stallard [1998]).

There are generally two Bayesian methods in determining sample size in a phase II trial. One such method is analogous to the frequentist approach where cut-offs are specified at the stage of design and the analysis is based on the Bayesian method before undertaking any decision, that is, to proceed to a definitive phase III trial or to abandon the development of the new drug. Thall and Simon [1994] presented a design and analysis of phase II clinical trials that is based on this method. The formulation of the model requires prior information of the response rate of the standard and new treatment, and a minimum ($n_{\min}$) and maximum ($n_{\max}$) total number of patients to be recruited.

In their design, patients are recruited sequentially and the total number of successes up to the i-th patient ($i = 1, 2, \ldots, n_{\max}$) is added up. Let X be the random variable of the number of successes, and Θ_S and Θ_E be the parameters of the response rates of standard and new treatments, respectively. If the observed cumulative number of successes X_i is greater than or equal to the upper bound U_i then the trial will terminate and the new treatment is declared promising. If X_i is less than or equal to the lower bound L_i then the trial will terminate and the new treatment is declared nonpromising. If $L_i < X_i < U_i$ then the phase II trial will continue by recruiting another patient. If however, at $i = n_{\max}$ and $L_i < X_i < U_i$ then the trial is concluded as inconclusive.

The upper and lower bounds are integers and are obtained from the posterior probability $\Pr(\Theta_E > \Theta_S + \delta_0 | X_i)$,

$$\begin{aligned} U_i &= \text{the smallest integer such that } \Pr(\Theta_E > \Theta_S + \delta_0 | X_i) \geq p_U \\ L_i &= \text{the largest integer such that } \Pr(\Theta_E > \Theta_S + \delta_0 | X_i) \leq p_L, \end{aligned}$$

where p_U and p_L are predetermined probabilities. The p_U should be preferably large, that is, between 0.95 and 0.99 and the p_L should be preferably small, $0.01 - 0.05$. The probabilities p_U and p_L are analogous to the power and type I error to the classical frequentist approach, respectively.

An extension of this method is one example from Tan and Machin [2002]. They proposed two Bayesian two-stage designs that resemble Simon's two-stage design but with similar Bayesian method as presented above. The first of such designs is known as a single threshold design (STD) and the second design is dual threshold design (DTD). In these models, instead of monitoring patients continuously, a group of patients is recruited to the first stage and the decision whether to continue the trial to recruit more patients to the stage two or to terminate the trial early and declare the new treatment is not promising is based on the posterior probability of the true response rate.

In the STD, supposed that λ_1 and λ_2 $(0 < \lambda_1 < \lambda_2 < 1)$ are the minimum thresholds at the interim stage and at the end of trial, respectively, that the true response rate is greater than the targeted response rate p_A. The hypothetical data from first and second stage of the design are used to determine an optimal sample size from stage one, n_1, and an optimal total sample size, n. The algorithm is used to search for the optimal n_1 and n under the constraints that the probability that the true response rate is greater than p_A based on the hypothetical data of stage one, and on the hypothetical data of stage one and two is at least λ_1 and λ_2, respectively.

The decision at the interim stage is based on the actual data from n_1 patients where the posterior probability that the true response rate is greater than p_A is obtained and if the probability is less than λ_1 then the phase II trial will cease and the drug is not recommended for further testing. Otherwise, the trial will continue to recruit the remaining $n - n_1$ patients into the stage two and at the end of the trial, the posterior probability that the true response rate is greater than p_A is obtained from all n patients. If the posterior probability is less than λ_2 then there is no strong justification to recommend the drug for definitive phase III trial. Otherwise, the drug is recommended to proceed to phase III trial.

The DTD is slightly different than the STD on the basis that the optimal sample size for stage one, n_1 is now determined by the probability that the true response rate is less than the control arm response rate p_0. The constraint is that the probability of the true response rate is less than p_0 is at least λ_1. The rationale is that if the posterior probability is greater than λ_1 then there is a high probability that the drug will be below the control arm response rate p_0, and as such, the drug will not be recommended for further testing and the trial will be abandoned. If however, the posterior probability is at least λ_1 then the remaining $n-n_1$ patients will be recruited to stage two and the analysis and decision at the end of the trial are the same as those in STD. Another difference between STD and DTD is that in DTD, λ_1 need not to be less than λ_2.

Another Bayesian method used in the design of phase II clinical trials is based on a decision theoretic approach where a loss or utility function in treating the patients in the phase II trial and the action taken at the end of trial is explicitly specified. The objective of this method is to optimise either the loss or the utility function. Some of the earlier works are by Sylvester [1988] and Brunier and Whitehead [1994].

In the design presented by Sylvester [1988] the loss function is composed of the loss in treating patients in the phase II trial and the action at the end of the trial. In conducting the phase II trial, the cost difference between "a patient who does not respond to the new treatment and a patient who does respond" is considered. At the end of the trial, there are two possible actions: (1) accept the drug and (2) reject the drug. In the event of accepting the drug, that is, the drug is recommended for further testing, the number of patients to be treated with the new treatment in phase III trials is considered in the loss function. The patient horizon which is "the average number of patients who are treated with an effective new drug after completion of the phase II trial before a second new drug which is at least as good is found" is also considered in the loss function.

Suppose that $d(x)$ is the action taken based on the observed data, then the loss function, $L(p, d(x))$, is the loss incurred upon taking action $d(x)$ when the true response rate is p. The form of the loss function is presented in Sylvester's paper. The risk function is the expectation of the loss function $L(p, d(x))$ and the optimal sample size is determined by minimising the risk function over a range of values of p.

In another design proposed by Brunier and Whitehead [1994], the model is based on the formulation of the cost of conducting an ineffective treatment (considered as a loss) and the expected gain if the treatment is found to be effective, and also the loss of rejecting an effective treatment. The gain and loss are fixed relative to each other. Similar to Sylvester's model, Brunier and Whitehead considered the number of patients who will be treated with the new treatment in phase III (if it shows to be promising in phase II) and if the new treatment is given to all future patients till a successor is found. One key assumption to the model is that the design of the phase III clinical trial is based on the conventional frequentist approach.

The response rate of the standard treatment in the phase II trial is assumed to be known whereas the probability of success of the new treatment is assumed to be unknown. Thus, a prior distribution is assumed for the new treatment. The optimal phase II design is obtained by maximising the expected utility function which is evaluated from the expected number of patients to be treated in phase II and the expected number of patients to be benefited from the new treatment.

3.3 A series of phase II trials

Thus far, all the designs described consider each phase II trial individually. In their review of statistical design and analysis of phase II trials, Mariani and Marubini [1996] described that there are two practical scenarios influencing the statistical frameworks. The first scenario is when patients are plentiful with a certain disease and new treatments are limited. In another scenario, the development of new therapies increases relatively faster than the recruitment of eligible patients. In the latter scenario it becomes more difficult to try all new therapies even with small sample sizes and therefore, there is a need to identify the promising ones effectively to be put into larger trials (see Whitehead [1985] and Yao et al. [1996]).

Some authors have proposed to consider a series of phase II trials as a "single trial" and the objective is to identify a promising trial in the quickest time. This is achieved by considering the relationship between the time of conducting the study and the number of patients for each trial: when the number of patients for each trial is optimised, the time to identify a promising treatment is also optimised. When finally a trial is declared promising, the "single trial" has achieved its goal and another "single trial" of a series of clinical trial can commence.

In Fred Hutchinson Cancer Research Center in Seattle the success of bone marrow transplantation as a treatment for the cancer of the blood has led to an increase in the number of trials searching for an ideal combination of preparative and follow-up procedures. The number of patients suitable for bone marrow transplantation is very small. This led to a model proposed by Whitehead [1985] (see also a qualitative discussion by Whitehead [1986]) which is suitable for rare diseases where the number of patients available is limited compared to the number of novel therapies waiting for trial.

One of the main assumption for Whitehead's approach is that the total number of patients available for study is considered known and fixed. Let the total number of patients be N. This number is usually a projection of the patient population eligible for trials for the next few months or years. Supposed that n patients are assigned to each of the distinct trial then there are a total of $M = N/n$ trials. Although not all the new treatments will be available for testing simultaneously, the total number of treatments can be projected based on the current development plan.

His model is to find the optimum number of trials to be studied and preferably all trials to run concurrently. The patients eligible for the study should be randomised to each treatment. If only one trial can be accepted for proof-of-concept phase III trial, then a selection method is used to estimate the best treatment after the analyses from each trial are available.

Let p_i be the probability of success for the i-th trial, $i = 1, 2, \ldots, M$, and assume they are independent random variables with a prior density of $g(p)$. Let $p_{[1]}$ denote the p_i from the most promising treatment based on the phase II data, that is the treatment selected for further testing in larger phase III trial. The optimal number of trials to be tested, M^*, is obtained by maximising the expected probability of success, $\mathrm{E}(p_{[1]})$, subject to the constraint of $N = nM$ where n and M have to be integers.

Another example of the design for a series of trials is motivated by vaccination studies in Memorial Sloan-Kettering Cancer Center (MSKCC) where there are many vaccines waiting to be put on trial. Yao et al. [1996] proposed a model to optimise the number of patients needed to find an effective treatment. Although the motivation behind Yao, Begg and Livingston's (hereafter known as YBL) model was based on vaccination, the term treatment will be used to be consistent with terms used throughout this report.

The most noticeable difference between YBL's proposed model from Whitehead's is that the total number of patients is not fixed. However, there is a more important difference between YBL's and Whitehead's. In YBL's model, the optimal sample size is determined under the constraints of some probability errors whereas in Whitehead's, as discussed above, the determination of optimal sample size is through the search of a maximum expected success probability in the selected treatment. Further details of the design of YBL's model is discussed below.

The main objective of YBL's model is to minimise the total number of patients, N. Each new treatments will be tested one-by-one till one is declared promising. Suppose that the M-th trial is the first trial to be declared promising then M is a random variable that follows a geometric distribution with $\Pr(M = m)$ as the probability of success.

According to YBL's model, let X_i be the number of successful responses in trial i for $i = 1, 2, \ldots$, and let p_i be the probability of a positive response for this vaccine. The parameters p_i are considered to be independent and identically distributed random variables following an underlying distribution $g(p)$. The aim of the vaccine screening is to find an effective vaccine with a probability of response $p_i \geq p_A$ where p_A is the "target" response rate. Thus, in each trial, the hypothesis testing is written formally as $H_0 : p_i < p_A$ against $H_1 : p_i \geq p_A$.

Suppose now that each trial evaluates n patients and it is declared as promising if the number of observed positive responses is greater than a critical value, k. Assuming that all the n patients are independent and if the first successful trial is called at the M-th trial then, $X_i|p_i \sim \text{binomial}(p_i, n) \quad i = 1, 2, \ldots, M$.

There are four possible outcomes from each trial: (1) do not reject H_0 and H_0 is true, (2) reject H_0 and H_0 is true, (3) do not reject H_0 and H_1 is true, and (4) reject H_0 and H_1 is true. The probabilities of each outcome is denoted respectively by,

$$\begin{aligned}
p_{--} = \Pr(C = 0, p_i < p_A) &= \int_{[0,p_A)} \Pr(X_i \le k|p)\,\mathrm{d}g(p) \\
&= \int_{[0,p_A)} \sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x}\,\mathrm{d}g(p), \\
p_{+-} = \Pr(C = 1, p_i < p_A) &= \int_{[0,p_A)} \Pr(X_i > k|p)\,\mathrm{d}g(p) \\
&= \int_{[0,p_A)} \sum_{x=k+1}^{n} \binom{n}{x} p^x (1-p)^{n-x}\,\mathrm{d}g(p), \\
p_{-+} = \Pr(C = 0, p_i \ge p_A) &= \int_{[p_A,1]} \Pr(X_i \le k|p)\,\mathrm{d}g(p) \\
&= \int_{[p_A,1]} \sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x}\,\mathrm{d}g(p), \\
p_{++} = \Pr(C = 1, p_i \ge p_A) &= \int_{[p_A,1]} \Pr(X_i > k|p)\,\mathrm{d}g(p) \\
&= \int_{[p_A,1]} \sum_{x=k+1}^{n} \binom{n}{x} p^x (1-p)^{n-x}\,\mathrm{d}g(p), \qquad (3.3)
\end{aligned}$$

where $C = 0$ denotes that the vaccine is declared to be nonpromising and $C = 1$ denotes that the vaccine is declared to be promising. Note that the error probabilities (3.3) are functions of n and k, and also depend on the distribution of $p, g(p)$. From these notations, the probability of a vaccine being declared promising can be written as $\Pr(M = m) = \Pr(X_i > k) = p_{+-} + p_{++} = p_{+\cdot}$.

There are two possible errors that could be made in the conclusion in a series of vaccine screening: (1) accepting a nonpromising treatment and (2) rejecting one or more promising vaccines. The authors showed that the probabilities of these errors are, respectively,

$$\begin{aligned}\alpha_1 &= \sum_{m=1}^{\infty} \Pr(X_1 \leq k, X_2 \leq k, \ldots, X_{m-1} \leq k, X_m > k, p_m < p_A) \\ &= \frac{p_{+-}}{p_{+\cdot}}, \qquad (3.4)\end{aligned}$$

and

$$\begin{aligned}\alpha_2 &= \sum_{m=1}^{\infty} \Pr(X_1 \leq k, X_2 \leq k, \ldots, X_{m-1} \leq k, X_m > k, \bar{g}_m) \\ &= \frac{p_{-+}}{p_{+\cdot} + p_{-+}}, \qquad (3.5)\end{aligned}$$

where

$$g_m = \{p_1 < p_A, p_2 < p_A, \ldots, p_{m-1} < p_A\}$$

and $\bar{g}_m$ is the complementary set to g_m. The optimal sample size (n^*) and cut-off (k^*) are obtained from a search algorithm by constraining the two posterior error probabilities; $\alpha_1 < e_1$ and $\alpha_2 < e_2$ where e_1 and e_2 are predefined maximum tolerable error rates.

Suppose that instead of controlling the false positive error rate for the whole series, the error rate is examined in the individual level of each trial. The posterior probability of a drug being nonpromising given that there are more than k positive responses observed is equal to $\Pr(p_i < p_A|C = 1) = \Pr(p_i < p_A|X_i > k)$. By some simple manipulation this is,

$$\Pr(p_i < p_A|X_i > k) = \frac{\Pr(p_i < p_A, X_i > k)}{\Pr(X_i > k)} = \frac{p_{+-}}{p_{+\cdot}}$$

which is the same as the definition of α_1. It shows that by controlling the posterior error rate at each trial level, the error rate of false positive of the whole series of trial is also controlled.

Although the false positive rate α_1 is incidentally easily interpreted in the levels of individual trial and whole series of trials, the false negative rate α_2 is not so straightforward. For example, in a study given by Leung and Wang [2001] the "target" response rate is $p_A = 0.2$. The error rates α_1 and α_2 are set to the maximum of 0.1 and 0.3, respectively. The average number of treatments to be tested is, $\mathrm{E}(M) = 6.2$, that is, there is an average of 5.2 rejected treatments before an effective one is found. The interpretation of α_2 is that the probability that at least one out of the 5.2 expected rejected treatments is promising is 0.3.

If however, the α_1 is now set at a maximum 0.2 while α_2 is maintained at 0.3, then the average number of rejected treatments is now 4.7 (that is, $\mathrm{E}(M) = 5.7$). The probability that at least one of the rejected treatments being promising is now based on a smaller average. The probability of rejecting any given promising treatments is now more serious than when it is based on a larger average. Thus, this underlines that the interpretation of α_2 depends on the expected number of treatments.

Building on the work by YBL, Leung and Wang [2001] extended the model by introducing a false negative of each individual trial. By definition, the probability of rejecting a promising vaccine is,

$$\alpha_2^* = \Pr(p_i \geq p_A | C = 0) = \frac{p_{-+}}{p_{-+} + p_{--}}.$$

By drawing on a relationship between frequentist error rates and Bayesian posterior error rates given by Lee and Zelen [2000] where both null and alternative hypotheses are simple hypotheses, the relationship between α_2 and α_2^* is,

$$\alpha_2^* = \frac{\alpha_2 \pi}{1 - (1 - \alpha_2)(\alpha_1 + \pi)}, \tag{3.6}$$

where π is the prior probability of being promising, that is, $\Pr(p_i = p_A) = \pi$.

The interpretation of α_2^* is clearer than that of α_2. Using the earlier example, for a pair of $(\alpha_1, \alpha_2) = (0.1, 0.3)$, and prior probability of $\pi = 0.217$, then $\alpha_2^* = 0.084$. If four vaccines were rejected before the fifth vaccine is declared promising, then the probability of a false negative for each rejected treatment is 8.4%. On the other hand, an $\alpha_2 = 0.3$ concludes that at least one out of the four rejected treatment is promising is 0.3.

The use of α_2 is appealing in the design of a series of clinical trials because the "overall" error rate can be controlled. However, the "practicality" of α_2 is limited in the interpretation of each rejected vaccine. Nevertheless, from the relationship shown in (3.6) it is easy to compute for α_2^* once α_2 has been specified and vice versa. It is then the choice of the investigator to decide whether to control for α_2 or α_2^* in the design.

Most of the designs discussed thus far, be it under the frameworks of classical frequentist or Bayesian methodology, estimated the optimal sample size by controlling two error rates; type I and II errors under the frequentist approach, and posterior probabilities equivalent to type I and II errors under the Bayesian approach. The designs that are not constrained by the error rates are those based on the decision theoretic approach (Sylvester [1988] and Brunier and Whitehead [1994]), and the model proposed by Whitehead [1985] which is to maximise the expected probability of the most promising treatment. In this report, the objective is to consider the practical scenario where the development of new therapies increases relatively faster than the recruitment of eligible patients. We set out to consider a design for a series of phase II clinical trials by controlling only one error, namely, the type I error and minimise the type II error which is equivalently in maximising the power of the trial. This hybrid design is proposed in the next chapter.

Chapter 4

Design of a series of clinical trials

In reality, a pharmaceutical company will have a number of drugs in development and waiting to be put on trial. Resources such as patients and money are essentially finite and limited. Therefore, a decision made on each clinical trial will subsequently affect the planning and development of other trials. It is then necessary to consider each clinical trial as part of a whole series of trials.

The objective of this project is to optimise the number of patients needed for each individual trial by maximising the power of each trial which is considered as part of a series of trials while maintaining a maximally accepted type I error at level α. Some of the methodology of this project extends works developed by other authors. The methodology employed in this project is a hybrid of frequentist and Bayesian methods where the traditional analysis at the end of the trial is based on the conventional frequentist hypothesis testing and the Bayesian method is used to maximise the power of a trial.

Although in practice, the number of patients per trial, n, is an integer, we will in this thesis as an idealisation assume n to be continuous. The advantage of considering n as a continuous variable is that it permits the maximisation of the power by differentiation.

4.1 Assurance

Similar to Whitehead's approach [Whitehead, 1985] assume that the total number of patients (N) is known and fixed. This number as described by Whitehead is usually a projection of the patient population eligible for the trial. In a series of sequential trials with n observations in each trial suppose that X_i is the sample mean of n observations from trial $i \quad (i = 1, 2, \ldots, M)$ and assume that it is normally distributed with mean θ_i and known variance σ^2/n. That is, $X_i \sim \mathrm{N}(\theta_i, \sigma^2/n)$. Consider now a simple two-sided hypothesis test

$$H_0 : \theta_i = \theta_0 \quad \text{vs.} \quad H_1 : \theta_i \neq \theta_0.$$

From the Neyman-Pearson lemma, a critical value k is chosen so that the test statistic rejects the null hypothesis at the desired level of α. That is, $\Pr(|X_i| > k) = \alpha/2$ if H_0 is true. Recall that the normal distribution is symmetric about its mean and the area of one end of the tail is the same as the area on the other end of the tail. Thus, to find the probability that $|X_i|$ is greater than k, it is sufficient to calculate $\Pr(X_i > k)$ and multiply by 2.

This gives

$$\begin{aligned} & \Pr(X_i > k) &= \alpha/2 \\ \Leftrightarrow\quad & \Pr\left(\tfrac{X_i-\theta_0}{\sqrt{\sigma^2/n}} > \tfrac{k-\theta_0}{\sqrt{\sigma^2/n}}\right) &= \alpha/2 \\ \Leftrightarrow\quad & 1-\Phi\left(\tfrac{k-\theta_0}{\sqrt{\sigma^2/n}}\right) &= \alpha/2 \\ \Leftrightarrow\quad & \tfrac{k-\theta_0}{\sqrt{\sigma^2/n}} &= z_{1-\alpha/2} \\ \Leftrightarrow\quad & k &= z_{1-\alpha/2}\sqrt{\sigma^2/n}+\theta_0 \end{aligned} \tag{4.1}$$

where $\Phi(\cdot)$ is the cumulative function of a standard normal distribution and $z(x)$ is the upper $100(1-x)\%$ point of the standard normal distribution.

We next consider the power of the trial to reject H_0 under $H_1 : \theta_i = \theta_A$ with θ_A assumed to be greater than θ_0. The power of a single trial is equal to

$$\begin{aligned} 1-\beta &= \Pr(X_i > k|\theta_i = \theta_A) \\ &= \Pr\left(Z > z_{1-\alpha/2} - \left(\frac{\theta_A-\theta_0}{\sqrt{\sigma^2/n}}\right)\right) \\ &= 1-\Phi\left(z_{1-\alpha/2} - \left(\frac{\theta_A-\theta_0}{\sqrt{\sigma^2/n}}\right)\right). \end{aligned} \tag{4.2}$$

The assumption that led to formula (4.2) is that θ_i is fixed. Suppose now that θ_i is random and follows a normal distribution with known mean μ and variance τ^2. The probability density function for θ_i is

$$g(\theta_i) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-(\theta_i-\mu)^2/(2\tau^2)}.$$

The "average" power over all the possible values of θ_i according to the prior belief which is called the *assurance* [O'Hagan and Stevens, 2001], is given by,

$$
\begin{aligned}
A &= \int_{-\infty}^{\infty} \left[1 - \Phi\left(z_{1-\alpha/2} - \sqrt{\frac{n}{\sigma^2}}(\theta_i - \theta_0)\right)\right] \frac{1}{\tau}\varphi\left(\frac{\theta_i - \mu}{\tau}\right) \mathrm{d}\theta_i \\
&= 1 - \int_{-\infty}^{\infty} \Phi\left(z_{1-\alpha/2} - \sqrt{\frac{n}{\sigma^2}}(\theta_i - \theta_0)\right) \frac{1}{\tau}\varphi\left(\frac{\theta_i - \mu}{\tau}\right) \mathrm{d}\theta_i \\
&= 1 - \int_{\theta_i=-\infty}^{\infty} \int_{x_i=-\infty}^{z_{1-\alpha/2}} \frac{1}{2\pi\tau} \exp\left[-\frac{1}{2}\left(\left(x - \sqrt{\frac{n}{\sigma^2}}(\theta_i - \theta_0)\right)^2 + \right.\right. \\
&\qquad \left.\left. \left(\frac{\theta_i - \mu}{\tau}\right)^2\right)\right] \mathrm{d}x_i \, \mathrm{d}\theta_i.
\end{aligned}
$$

For ease of notation, denote $\frac{\sigma^2\mu + \sqrt{n}\sigma\tau^2 x_i + n\tau^2\theta_0}{\sigma^2 + n\tau^2}$ by λ and $\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}$ by ϕ^2, then the above expression is simplified to,

$$
\begin{aligned}
A = 1 &- \int_{\theta_i=-\infty}^{\infty} \int_{x_i=-\infty}^{z_{1-\alpha/2}} \frac{1}{2\pi\tau} \exp\left[-\frac{1}{2}\left(\left(\frac{\theta_i - \lambda}{\phi}\right)^2 + \right.\right. \\
&\qquad \left.\left. \frac{(\sigma x_i - \sqrt{n}(\mu - \theta_0))^2}{\sigma^2 + n\tau^2}\right)\right] \mathrm{d}x_i \, \mathrm{d}\theta_i \\
= 1 &- \int_{x_i=-\infty}^{z_{1-\alpha/2}} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{(x_i - \sqrt{n/\sigma^2}(\mu - \theta_0))^2}{2(1 + n\tau^2/\sigma^2)}\right] \cdot \\
&\qquad \int_{\theta_i=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(\theta_i - \lambda)^2}{2\phi^2}\right] \mathrm{d}\theta_i \, \mathrm{d}x_i \\
= 1 &- \int_{-\infty}^{z_{1-\alpha/2}} \frac{1}{\sqrt{2\pi(1 + n\tau^2/\sigma^2)}} \exp\left[-\frac{(x_i - \sqrt{n/\sigma^2}(\mu - \theta_0))^2}{2(1 + n\tau^2/\sigma^2)}\right] \mathrm{d}x_i \\
= 1 &- \Phi\left(\frac{z_{1-\alpha/2} - \sqrt{n/\sigma^2}(\mu - \theta_0)}{\sqrt{1 + n\tau^2/\sigma^2}}\right).
\end{aligned}
$$

Note that the assurance can also be obtained directly from the marginal distribution of X_i. From its likelihood function, $f(x_i|\theta_i) = \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-(x_i-\theta_i)^2/(2\sigma^2/n)}$, the joint distribution of X_i and θ_i is

$$\begin{aligned}
h(x_i, \theta_i) &= f(x_i|\theta_i)g(\theta_i) \\
&= \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{ -\frac{(x_i-\theta_i)^2}{2\sigma^2/n} \right\} \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{ -\frac{(\theta_i-\mu)^2}{2\tau^2} \right\} \\
&= \frac{1}{2\pi\tau\sigma/\sqrt{n}} \exp\left\{ -\frac{1}{2}\left[\left(\frac{x_i-\theta_i}{\sigma/\sqrt{n}} \right)^2 + \left(\frac{\theta_i-\mu}{\tau} \right)^2 \right] \right\}. \qquad (4.3)
\end{aligned}$$

The marginal distribution of X_i is obtained by integrating $h(x_i, \theta_i)$ over θ_i,

$$\begin{aligned}
f_{X_i}(x_i) &= \int_{-\infty}^{\infty} f(x_i|\theta_i)g(\theta_i)\,\mathrm{d}\theta_i \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\tau\sigma/\sqrt{n}} \exp\left\{ -\frac{1}{2}\left[\left(\frac{x_i-\theta_i}{\sigma/\sqrt{n}} \right)^2 + \left(\frac{\theta_i-\mu}{\tau} \right)^2 \right] \right\} \mathrm{d}\theta_i \\
&= \frac{\sqrt{n}}{\sigma\tau\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left[\frac{n}{n\tau^2+\sigma^2}(x_i-\mu)^2 \right] \right\} \cdot \\
&\qquad \int_{-\infty}^{\infty} -\frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\theta_i - \frac{n\tau^2 x_i + \sigma^2\mu}{n\tau^2+\sigma^2} \right)^2 \left(\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2} \right) \right\} \mathrm{d}\theta_i \\
&= \frac{\sqrt{n}}{\sigma\tau\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left[\frac{n}{n\tau^2+\sigma^2}(x_i-\mu)^2 \right] \right\} \cdot \left(\frac{\sigma\tau}{\sqrt{n\tau^2+\sigma^2}} \right) \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2/n)}} \exp\left\{ -\frac{(x_i-\mu)^2}{2(\tau^2+\sigma^2/n)} \right\}.
\end{aligned}$$

The marginal distribution of the random variable, X_i, has the form of a normal distribution with mean μ and variance $(\tau^2 + \sigma^2/n)$. Subsequently, based on (4.1) and (4.2), the assurance is,

$$\begin{aligned} A &= \Pr(X_i > k|\theta_i \sim \mathrm{N}(\mu, \tau^2)) \\ &= \Pr(X_i > z_{1-\alpha/2}\sqrt{\sigma^2/n} + \theta_0|\theta_i \sim \mathrm{N}(\mu, \tau^2)) \\ &= 1 - \Phi\left(\frac{z_{1-\alpha/2}\sqrt{\sigma^2/n} + \theta_0 - \mu}{\sqrt{\tau^2 + \sigma^2/n}}\right) \\ &= 1 - \Phi\left(\frac{z_{1-\alpha/2} - \sqrt{n/\sigma^2}(\mu - \theta_0)}{\sqrt{1 + n\tau^2/\sigma^2}}\right), \end{aligned}$$

as obtained above. Examining the mathematical property of the assurance, as $n \to 0$, $A = 1 - \Phi(z_{1-\alpha/2}) = \alpha/2$, the specified one-sided type I error rate. This suggests that there is a minimum power, albeit very small, that can be attained when the sample size goes to 0, which means that we may still benefit from a successful trial without even starting a trial! Whereas, as $n \to \infty$, $A \approx 1 - \Phi(\frac{z_{1-\alpha/2} - \sqrt{n/\sigma^2}(\mu-\theta_0)}{\sqrt{n\tau^2/\sigma^2}}) \to \Phi(\frac{\mu-\theta_0}{\tau})$, the prior probability that $\theta_i > \theta_0$. Therefore, if the prior belief is positive, the assurance will be high and if the prior is negative, the assurance will be low.

4.1.1 Model 1: Maximisation of assurance

The assurance can be interpreted as the average probability of rejecting the null hypothesis over all possible values of the parameter of interest that is based on the prior distribution. As mentioned earlier, in a series of clinical trials where the total number of patients has been fixed as N and assuming that each individual trial requires n number of patients, the total number of trials to be tried in the series is simply $M = N/n$. Let $\tilde{M}$ be the number of trials that reject H_0. The first design we are proposing is to find the optimal number of patients per trial, n^*, that maximises the expected number of trials that reject H_0, $\mathrm{E}(\tilde{M})$. Recall that the normal density is symmetric about its mean and for the following designs, we will continue to work on only one of the normal density tails and thus, the first assumption is that $\mu - \theta_0 > 0$. The expected number of trials that reject H_0,

$$\begin{aligned}\mathrm{E}(\tilde{M}) &= MA \\ &= \frac{N}{n}\biggl(1 - \Phi\biggl(\frac{z_{1-\alpha/2} - \sqrt{n/\sigma^2}(\mu - \theta_0)}{\sqrt{1 + n\tau^2/\sigma^2}}\biggr)\biggr). \end{aligned} \tag{4.4}$$

As $nM = N$ is fixed, $\mathrm{E}(\tilde{M})$ can be considered as a function of n or M. As shown earlier, as $n \to \infty$ (note though n can only be as great as N because the total number of patients is fixed but suppose that in an ideal situation where $N \to \infty$, then $n \to \infty$), $N/n \to 0$ and the assurance, as shown earlier, is bounded by the prior probability that $\theta_i > \theta_0$. Therefore, $\mathrm{E}(\tilde{M}) \to 0$. On the contrary, as $n \to 0$, the assurance will go towards the fixed error rate, $\alpha/2$, that is, the second term of the $\mathrm{E}(\tilde{M})$ is bounded, whereas, the first term, $N/n \to \infty$. Therefore, $\mathrm{E}(\tilde{M}) \to \infty$. As there is no other value that is greater than $+\infty$, the global maximum is at $n = 0$. Thus, the optimal sample size, $n^* = 0$.

As an example of the property of $\mathrm{E}(\tilde{M})$, consider for an i-th trial where the planned analysis is set to test the null hypothesis of $H_0 : \theta_i = 0$ against the alternative, $H_1 : \theta_i \neq 0$, and we fixed the maximally accepted type I error at, $\alpha = 0.05$. Supposed that the sample mean X_i follows a normal distribution with mean θ_i and variance $1/n$ (that is, $\sigma = 1$) and the random parameter θ_i has a prior distribution that follows a normal distribution with

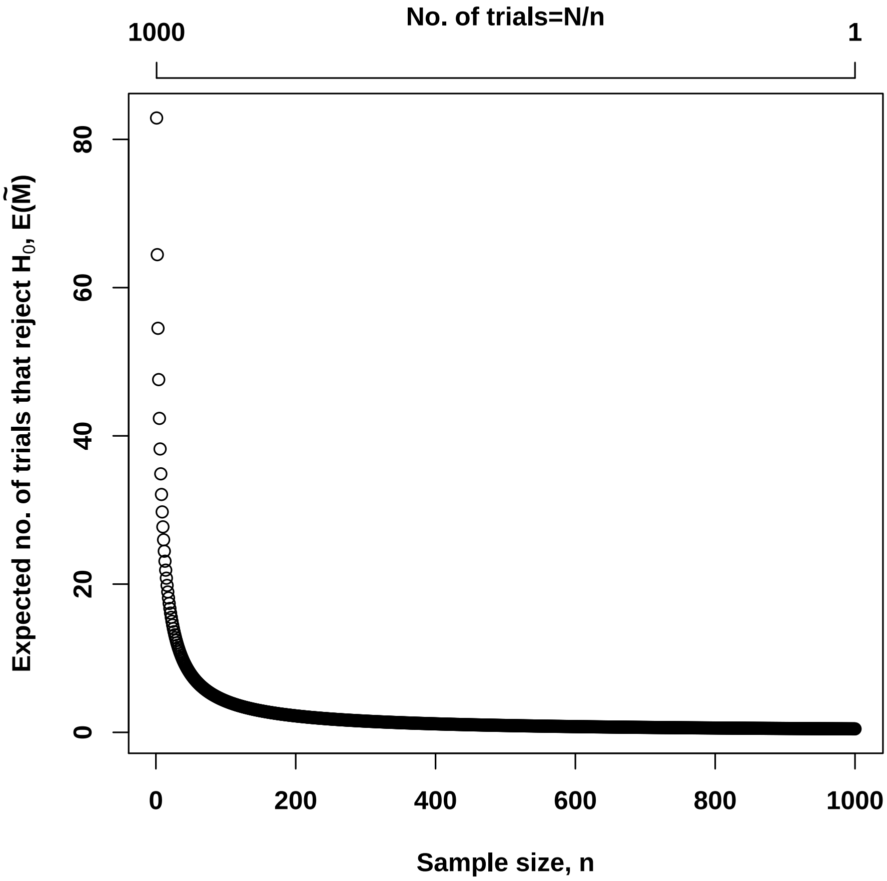


Figure 4.1: The expected number of trials that reject null hypothesis against the sample size for each individual trial. The hypothesis is $H_0 : \theta_i = 0$ vs $H_1 : \theta_i \neq 0$. Based on $X_i \sim \mathrm{N}(\theta_i, 1/n)$, $\theta_i \sim \mathrm{N}(0, 1)$, $N = 1000$, and $\alpha = 0.05$.

mean 0 and variance 1. Assuming that the projection of the total sample size is, $N = 1000$, Figure 4.1 shows the various values of $\mathrm{E}(\tilde{M})$ with different values of n.

Corresponding to the mathematical interpretation, the plot shows that to maximize the number of trials to be recommended for definitive phase III trials when they are showing efficacy, we should have many small individual trials. In fact, if n is fixed to have only integer value, each individual trial should have only one patient. By having only one patient per trial, naturally more drugs can be put on trial and consequently, increases the expected total number of trials that reject H_0 since there is positive probability of rejection of H_0 even with a very sample size.

It appears that $\mathrm{E}(\tilde{M})$ is a decreasing function of n. Due to the monotonic function, if there is a constraint on the minimum number of patients required for each trial, then that minimum n will give the highest expected number of trials that reject null hypothesis.

4.1.2 Model 2: Minimisation of the expected net loss

So far, we have ignored the start-up cost per trial and hence based on the preceding result, we "get something for nothing" and it could be this that leads to the very small optimal sample size. In practice, there is a start-up cost associated with each clinical trial. By having as many trials as the total number of patients available, the total start-up cost will be greatly inflated. The start-up cost could be the money or the time spent on planning, designing, submitting for ethics approval, and so on.

Suppose that one unit of value is assigned to each successful trial, then the total expected gain is $\mathrm{E}(\tilde{M})$. Let the start-up cost be fixed at f which is relative to the one unit of gain. The total start-up cost for all trials is thus, fM. The expected net loss is,

$$\mathrm{E}(F) = fM - \mathrm{E}(\tilde{M}). \tag{4.5}$$

Note that the total cost does not include the cost for each patient. This is because the total number of patients is known and fixed. Therefore, the total patients cost is a constant and will not affect the optimisation of the model.

The optimisation problem is to find an n that minimises the expected net loss, $\mathrm{E}(F)$. To do so, the expression,

$$\begin{aligned}\mathrm{E}(F) &= f\left(\frac{N}{n}\right) - \frac{N}{n}\left(1 - \Phi\left(\frac{z_{1-\alpha/2} - \sqrt{n/\sigma^2}(\mu - \theta_0)}{\sqrt{1 + n\tau^2/\sigma^2}}\right)\right) \\ &= \frac{N}{n}\left(f + \Phi\left(\frac{z_{1-\alpha/2} - \sqrt{n/\sigma^2}(\mu - \theta_0)}{\sqrt{1 + n\tau^2/\sigma^2}}\right) - 1\right)\end{aligned}$$

can be differentiated with respect to n, and subsequently from the equation $\mathrm{dE}(F)/\,\mathrm{d}n = 0$ we solve for n. However, due to the $\Phi(\cdot)$ term in the differentiated expression, this can only be solved numerically. The search for the optimal n^* that minimises the expected net loss, $\mathrm{E}(F)$, can alternatively be made by a direct computation of $\mathrm{E}(F)$ for a range of values for n (from 0.01 to N by an increment of 0.01). The n that corresponds to the smallest value of $\mathrm{E}(F)$ is then considered as the optimal n^*.

For an example on the characteristics of the design, again consider that in each trial the null hypothesis of $H_0 : \theta_i = 0$ is tested against the alternative, $H_1 : \theta_i \neq 0$, and α is set at 0.05. We have $X_i \sim \mathrm{N}(\theta_i, \sigma^2)$, and the prior distribution is $\theta_i \sim \mathrm{N}(\mu, \tau^2)$. Suppose that the total sample size is $N = 1000$, and the start-up cost is fixed at $f = 0.05$. Figure 4.2 shows that both fM and $\mathrm{E}(\tilde{M})$ are monotonic functions of n.

Figure 4.3(a) shows that the curve of the net loss, $\mathrm{E}(F)$, gets “narrower” as τ increases while σ is held constant ($\sigma = 1$). When τ gets larger, the variance of the prior belief is wider. Thus, smaller sample sizes are required to give information regarding the θ_i from each trial. In contrast to what has been observed from the increment of τ, Figure 4.3(b) shows that $E(F)$ is "stretched" as σ increases while τ is held constant ($\tau = 5$). This property conforms to the idea that as the standard deviation of the likelihood function is wider, larger sample size is required to deliver more information on the parameter of interest, θ_i.

Table 4.1 presents some of the optimal sample sizes (n^*) under different likelihood functions and prior distributions such that the effect sizes (μ/σ) are 0.2 ("small" effect size), 0.5 ("moderate" effect size) and 0.8 ("large" effect size). Suppose that the prior mean, $\mu = 1$, and the standard deviation of the likelihood function, $\sigma = 2$, (giving a moderate prior effect size of 0.5) and the standard deviation of the prior mean is $\tau = 1$, an optimal sample size $n^* = 2.37$ gives the smallest expected net loss, $\mathrm{E}(F) = -51.813$. From a total of 1000 patients and if 2.37 patients were recruited in each trial, the expected number of trials that will reject H_0 is $\mathrm{E}(\tilde{M}) = 72.91$ out of 421.941 trials that will be conducted. Note that when the parameters ($\mu = 1, \tau = 1, \sigma = 2$) is multiplied by 2 for example, ($\mu = 2, \tau = 2, \sigma = 4$), the n^* is the same, that is, $n^* = 2.37$. The optimal sample size n^* is robust towards the unit of measurement. For example, if the original measurement is in centimetres and it has now changed to inches, the n^* will remain the same regardless.

There are some situations where an optimal n^* cannot be found. As the fixed start-up cost, $f \to 0$, the optimal sample size, $n^* \to 0$. As shown from Figure 4.4, when f is very small the total cost for starting trials, fM, is also negligible (refer to (4.5)). Thus, the model is the same as Model 1 that has been proposed earlier, which in order to maximise the power of a series of trial is to conduct as many trials as possible with only one patient.

Table 4.2 shows the optimal sample sizes when f takes on various values between 0 and 1 while holding other parameters fixed. When $f = 0.01$, $n^* = 0.01$ which is the minimum value used in the direct search algorithm. By modifying the minimum value in the direct search algorithm n^* changes and takes on that minimum value (result not shown).

On the other hand, as $f \to 1$, the optimal sample size $n^* \to \infty$. Practically, n^* will take on the value of N as the total number of patients has been fixed before the trials begun. Referring back to the equation (4.5), as $f \to 1$, the total cost, fM, will always be greater than the expected gain and it is thus very expensive to start any trial. Henceforth, it is advisable to have only one trial with all the patients in it. In the example shown in Table 4.2, n^* takes on 1000 when the start-up cost is 0.99. In other scenarios when $f = 0.99$ and N varies with smaller and larger values than 1000, n^* is always equal to N (result not shown).

4.1.3 Model 3: Minimisation of the total cost

So far, our designs have a constraint of a fixed total sample size N. Suppose now that there are unlimited number of patients for a series of trial and consequently, the total number of trials including the first successful trial is also not fixed. This approach will be the same as that proposed by Yao et al. [1996]. In the formulation of Model 3, consider a series of phase II trials as one single trial. The objective of the series is to find the first trial that is successful and recommended for further testing in definitive phase III trial.

When the first successful trial is obtained, the series ends and another series of trials will start.

The total number of trials including the first successful trial is now a random variable which follows a geometric distribution. The prior probability that a trial that is declared successful is equal to A, the assurance as introduced earlier. The expected number of trials required to give one successful trial, including the first successful trial is $1/A$. Following on the notation in Model 2, let f be the start-up cost per trial, then the average total start-up cost for all trials up to and including the first successful trial is f/A.

There are n patients in each trial so that the expected total number of patients till a successful trial is found is $\mathrm{E}(n) = n/A$. As the total number of patients to be required is not fixed, the total cost needed to be spent on patients is not fixed either. Let c be the cost per patient which is also relative to the one unit of gain, then the expected total cost of patients will be $c\mathrm{E}(n) = cn/A$.

Therefore, the expected total cost that would be spent till a successful trial is declared is simply

$$\begin{aligned} \bar{C} &= f/A + c\mathrm{E}(n) \\ &= \frac{1}{A}(f + cn). \end{aligned} \tag{4.6}$$

The optimisation problem is to find an n that minimises the expected total cost, $\bar{C}$. Similar to Model 2, the expression in (4.6) can be differentiated with respect to n. However, due to the $\Phi(\cdot)$ term in the assurance, this can only be solved numerically. Figure 4.5 shows the three expected total costs of conducting a series of trials: the expected cost of patients ($c\mathrm{E}(n)$), the expected start-up cost (f/A), and the expected total cost ($\bar{C}$) against the sample size per trial (n). Clearly, from both equation (4.6) and Figure 4.5, the expected cost of patient increases linearly with respect to n whereas the expected start-up cost decreases exponentially with respect to n.

Table 4.3 shows some optimum sample sizes for each trial that minimise the expected total costs for the effect sizes (μ/σ) of 0.2, 0.5, and 0.8. The search algorithm used to produce the table begins with a very small value for n (0.01 in this example) and increases to 200 by 0.01. The maximum value of 200 was chosen because most phase II trials have less than 100 patients. An n is considered optimum if it corresponds to the smallest value of $\bar{C}$.

Suppose that the null hypothesis $H_0 : \theta_i = 0$ of each trial in a series is tested against an alternative $H_1 : \theta_i \neq 0$. Let the start-up cost per trial be $f = 0.02$, the cost per patient be $c = 0.001$. If the likelihood function $f(x_i|\theta_i)$ has a normal distribution with mean θ_i and standard deviation $\sigma = 5$, and the prior distribution is normally distributed with mean $\mu = 1$ and standard deviation $\tau = 2$, and setting the $\alpha = 0.05$, the optimal sample size per trial is $n^* = 19.22$ (Table 4.3). The expected total number of patients needed till the first successful trial is found is $\mathrm{E}(n) = 64.979$.

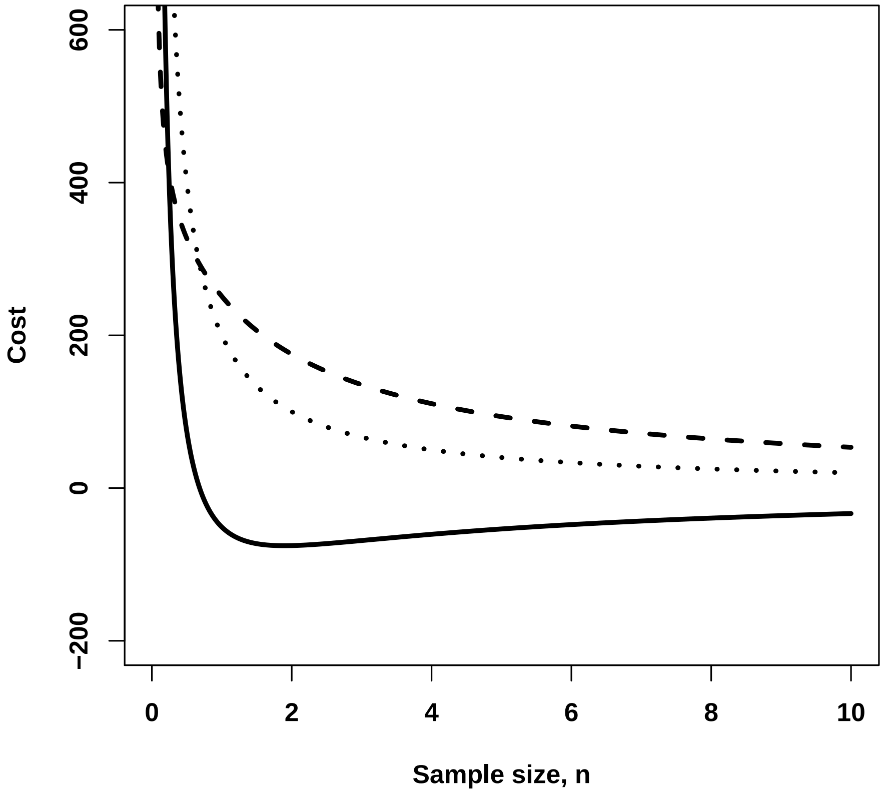
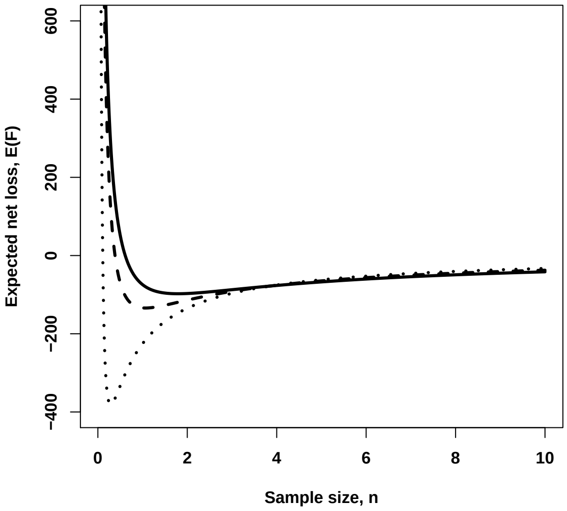
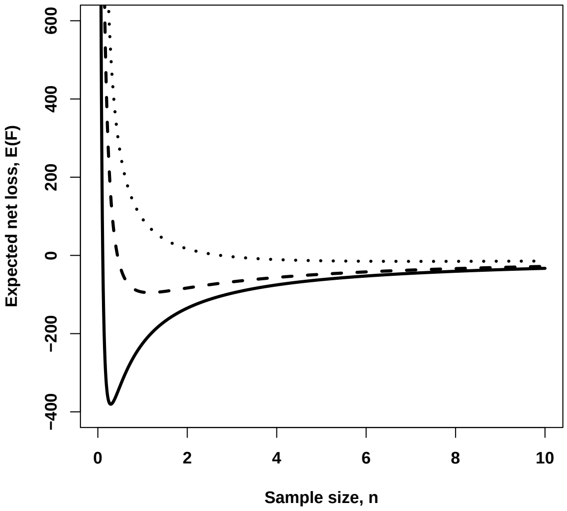


Figure 4.2: The expected net loss from conducting a series of trials, $\mathrm{E}(F)$, (solid), the total cost of trials, fM, (dotted) and the expected gain, $\mathrm{E}(\tilde{M})$, (dashed) against the sample size for each individual trial, n. The hypothesis is, $H_0 : \theta_i = 0$ vs. $H_1 : \theta_i \neq 0$. Based on $X_i \sim \mathrm{N}(\theta_i, 4)$ and $\theta_i \sim \mathrm{N}(1.5, 9)$, and $\alpha = 0.05$.



(a) $\mu = 1$ and $\sigma = 1$



(b) $\mu = 1$ and $\tau = 5$

Figure 4.3: The expected net loss, E(F), from conducting a series of trials against the sample size for each individual trial, n, for (a) τ of 1.25 (solid), 2.0 (dashed), and 5.0 (dotted) and (b) σ of 1.0 (solid), 2.0 (dashed), and 5.0 (dotted).

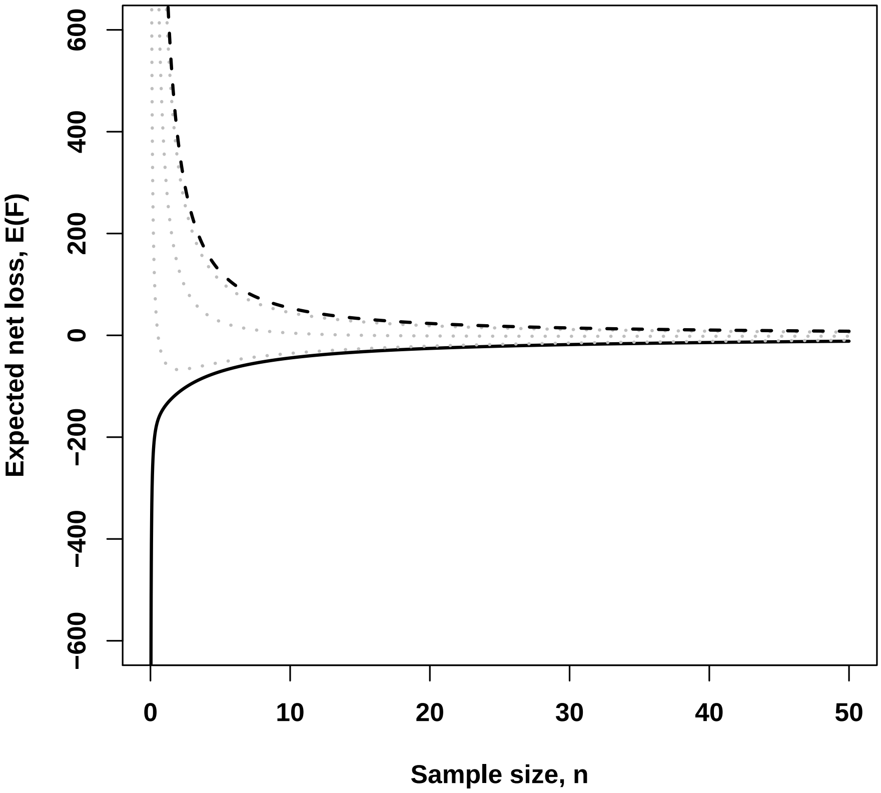


Figure 4.4: The expected net loss from conducting a series of trials against the sample size for each individual trial for f of 0.01 (black solid line), and 0.99 (black dashed line). The other grey dotted lines in between are various f values at 0.1, 0.5, and 0.9 as the curve changes from "narrow" to "spread out".

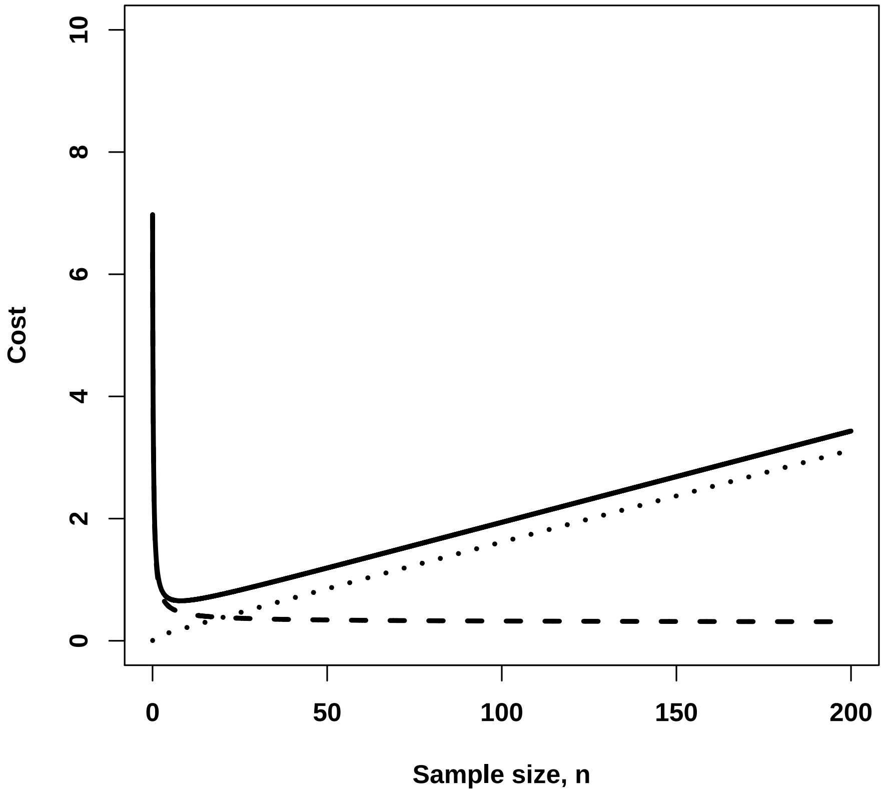


Figure 4.5: The expected total cost of a series of trials $\bar{C}$ (solid), the expected start-up cost f/A (dashed), and the expected cost of patients $c\mathrm{E}(n)$ (dotted) against the sample size for each individual trial n. The hypothesis is, $H_0 : \theta_i = 0$ vs. $H_1 : \theta_i \neq 0$. Based on $X_i \sim \mathrm{N}(\theta_i, 4)$ and $\theta_i \sim \mathrm{N}(1, 4)$, and $\alpha = 0.05$.

Table 4.1: Optimal sample sizes per trial, n^*, expected gain, $\mathrm{E}(\tilde{M})$, and expected net loss, $\mathrm{E}(F)$, under various likelihood functions and prior distributions. The total number of patients is fixed at $N = 1000$ and the start-up cost is $f = 0.05$. The hypothesis is, $H_0 : \theta_i = 0$ vs. $H_1 : \theta_i \neq 0$. Based on $X_i \sim \mathrm{N}(\theta_i, \sigma^2)$, $\theta_i \sim \mathrm{N}(\mu, \tau^2)$, and $\alpha = 0.05$.

Effect size	σ	μ	τ	n^*	$\mathrm{E}(\tilde{M})$	$\mathrm{E}(F)$
0.2	5	1	1	14.83	11.662	-8.29
	5	1	2	6.38	23.991	-16.154
	5	1	5	1.51	92.469	-59.356
	7.5	1.5	1	21.53	8.829	-6.507
	7.5	1.5	2	10.78	15.158	-10.52
	7.5	1.5	5	2.96	48.697	-31.805
	10	2	1	26.66	7.689	-5.813
	10	2	2	14.83	11.662	-8.29
	10	2	5	4.62	32.173	-21.35
0.5	2	1	1	2.37	72.91	-51.813
	2	1	2	1.02	149.983	-100.963
	2	1	5	0.24	579.307	-370.974
	3	1.5	1	3.44	55.201	-40.666
	3	1.5	2	1.73	94.649	-65.748
	3	1.5	5	0.47	305.162	-198.779
	4	2	1	4.27	48.043	-36.334
	4	2	2	2.37	72.91	-51.813
	4	2	5	0.74	201.006	-133.438
0.8	1.25	1	1	0.93	186.405	-132.642
	1.25	1	2	0.4	383.465	-258.465
	1.25	1	5	0.09	1504.754	-949.198
	1.875	1.5	1	1.35	141.141	-104.104
	1.875	1.5	2	0.67	242.94	-168.313
	1.875	1.5	5	0.19	771.955	-508.797
	2.5	2	1	1.67	122.954	-93.014
	2.5	2	2	0.93	186.405	-132.642
	2.5	2	5	0.29	514.014	-341.601

Table 4.2: Optimal sample sizes per trial, n^*, expected gain, $\mathrm{E}(\tilde{M})$, and expected net loss, $\mathrm{E}(F)$, under various start-up costs, f. The hypothesis is, $H_0 : \theta_i = 0$ vs. $H_1 : \theta_i \neq 0$. Based on $X_i \sim \mathrm{N}(\theta_i, 4)$, $\theta_i \sim \mathrm{N}(1, 4)$, $N = 1000$, and $\alpha = 0.05$.

f	n^*	$\mathrm{E}(\tilde{M})$	$\mathrm{E}(F)$
0.001	0.01†	2868.474	-2768.474
0.01	0.01†	2868.474	-1868.474
0.05	1.02	149.983	-100.963
0.1	2.04	116.397	-67.378
0.2	4.26	80.488	-33.54
0.5	34.69	16.301	-1.887
0.6	143.24	4.407	-0.218
0.9	1000	0.669	0.231
0.99	1000	0.669	0.321

† In the direct search algorithm a value of 0.01 was used as the minimum for n to start off the search. Hence, the minimum value that n^* can reach is 0.01.

Table 4.3: Optimal sample sizes per trial, n^*, expected total number of patients, $\mathrm{E}(n)$, and expected total cost, $\bar{C}$, under various start-up costs, f, and costs per patient, c. The hypothesis is, $H_0 : \theta_i = 0$ vs. $H_1 : \theta_i \neq 0$. Based on $X_i \sim \mathrm{N}(\theta_i, \sigma^2)$, $\theta_i \sim \mathrm{N}(\mu, \tau^2)$, and $\alpha = 0.05$.

Effect size	σ	μ	τ	c	f	n^*	$\mathrm{E}(n)$	$\bar{C}$
0.2	5	1	2	0.001	0.05	31.08	85.205	0.222
	5	1	2	0.002	0.05	21.61	69.088	0.298
	5	1	2	0.001	0.02	19.22	64.979	0.133
	7.5	1.5	2	0.001	0.05	42.17	113.833	0.249
	7.5	1.5	2	0.002	0.05	29.58	95.339	0.352
	7.5	1.5	2	0.001	0.02	26.34	90.545	0.159
	10	2	2	0.001	0.05	51.79	135.15	0.266
	10	2	2	0.002	0.05	36.53	115.587	0.389
	10	2	2	0.001	0.02	32.5	110.391	0.178
0.5	2	1	2	0.001	0.05	13.37	27.502	0.13
	2	1	2	0.002	0.05	9.14	20.585	0.154
	2	1	2	0.001	0.02	8.1	18.865	0.065
	3	1.5	2	0.001	0.05	17.37	33.615	0.13
	3	1.5	2	0.002	0.05	12.09	25.997	0.16
	3	1.5	2	0.001	0.02	10.77	24.081	0.069
	4	2	2	0.001	0.05	20.61	37.492	0.128
	4	2	2	0.002	0.05	14.56	29.693	0.161
	4	2	2	0.001	0.02	13.03	27.721	0.07
0.8	1.25	1	2	0.001	0.05	8.88	16.608	0.11
	1.25	1	2	0.002	0.05	6	12.001	0.124
	1.25	1	2	0.001	0.02	5.29	10.852	0.052
	1.875	1.5	2	0.001	0.05	11.25	19.491	0.106
	1.875	1.5	2	0.002	0.05	7.74	14.5	0.123
	1.875	1.5	2	0.001	0.02	6.87	13.253	0.052
	2.5	2	2	0.001	0.05	13.05	21.066	0.102
	2.5	2	2	0.002	0.05	9.13	16.035	0.12
	2.5	2	2	0.001	0.02	8.15	14.773	0.051

Chapter 5

Further work

The risk of drug research and development is increasing while the profit is decreasing. In 1996, Senn [1996] stressed that it is important to choose a portfolio of drugs in development as it may be an essential to the survival of a pharmaceutical company. Resources are finite. Thus, not all projects can be developed. In the clinical trials of drug development, the trials are arranged sequentially and not all development costs are paid up-front. The failure of a drug at any phase may give the opportunity to abandon the project and thus there is no commitment to fund the drug development any further. Julious and Swank [2005] also emphasized that when there are a number of drugs waiting for development, a formal decision analysis is essential especially in the "potential of fast failure".

Motivated by these ideas, we therefore plan to build on the current work by considering other aspects of clinical trials. The formulation will be based on the scenario where the supply of new therapies is high and the targeted patient population is limited. The following section will give some details on the extensions to be considered.

5.1 Extension

Some of the further works planned for the next two academic years are:

1. Tables 4.1 in Chapter 4 are based on a fairly large total number of patients, $N = 1000$. However, in rare diseases or diseases with small populations such as paediatric, geriatric and ethnic minority populations, it is unrealistic to expect so many patients eligible for trial for any given year. Thus, we will inspect qualitatively how Model 2 will behave when N is considerably smaller, that is, $N < 500$. Due to the much smaller total sample size, the number of trials that can be tested in a series will be less. A problem that we anticipate is, for example, suppose that $N = 250$ and the optimal sample size per trial is $n^* = 27$. At the end of the 8-th trial, a total of 216 patients would have been enrolled into the series of trials and there are only 34 patients left available for recruitment. If however, there is not any trial that rejects H_0 up to the 8-th trial should the investigator continue to recruit 27 patients to one more trial and hope that this trial will reject the null hypothesis or is there an alternative? We wish to examine further the anticipated problems and various possibilities.

2. The Model 2 proposed in Chapter 4 assumed that there must be at least one trials in a series of trial that would reject the null hypothesis. What if none of the trials reaches the minimum cut-off for the investigator to

declare that the new treatment is promising? A different type of design is necessary to incorporate such scenario.

3. Some of the choices of actions in a phase II trial are to (i) proceed to phase III trial with the new treatment, (ii) continue to another phase II trial with the same treatment, (iii) go back to phase I trial with the same treatment but perhaps with different dose or combinations or route of administration, etc, and (iv) cease the development. These actions with their corresponding loss and gain, and probability of success and failure could be included in the next design. Again a special case of this extension is to consider a limited N.

4. The Models 1, 2, and 3 are based on the assumptions that each individual trial has the same prior beliefs. However, different compounds under study may have slightly different probability of effectiveness. For example, some treatments are modifications of existing effective treatments or some treatments that are well established in other diseases but not in the population on the current trial or some treatments are quite new so that the prior information is obtained from animal or laboratory studies. It is of interest to investigate how the optimisation problem will be affected under different priors.

5. Depending on the approval from the industrial collaborator, Roche, the models described above will be fitted with real-life data for illustration and qualitative comparison.

5.2 Proposed headings for the thesis

The proposed headings for the thesis will be as followed:

Chapter 1: Background

Chapter 2: Design of a series of phase II clinical trials

Chapter 3: Design of a series of projects

Chapter 4: Application of proposed designs with real-life data

Chapter 5: Discussion

Chapter 6

Completed and planned training

List of completed training

The list below is the trainings completed via Warwick Graduate School Skills Programme in the year 2009.

Feb 21: Speed Reading

Feb 25: Writing a PhD Thesis - Science and Medicine

Mar 04: Thesis structure

Mar 11: Time Management and Self Motivation

Apr 14: 2 day Effective Researcher Training Course

Apr 24: Preparing for the Upgrade from Mphil to PhD

May 08: Presenting to an Academic Audience

May 22: Voice Training

Planned training

One of the planned trainings is an attachment to the statistical division of Roche for a few weeks.

The list below is some of the courses and trainings planned for the next two years. The courses are organised by Academy for PhD Training in Statistics (APTS), Royal Statistical Society (RSS) and the Postgraduate Statistics Centre of Lancaster University.

- Statistical inference
- Statistical computing
- Bayesian adaptive designs
- Presenting statistical data
- Ethics and statistics
- Scientific writing

Besides courses and trainings, at least two conferences are planned for the next two years:

- Research Students' Conference in Probability and Statistics, 12th-15th April 2010 in the University of Warwick.

- International Society for Clinical Biostatistics Conference, August 2011 in Ottawa Canada.

Tables 6.1 and 6.2 show some of the trainings and programmes planned for the next two academic years. The table will be updated when other relevant courses, trainings and conferences are available.

Table 6.1: Diary of 2009/2010

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Upgrading report	■	■										
Work on proposed extension (1)†			■	■	■	■						
Prepare for APTS week 1			■									
APTS week 1				■								
Prepare abstract for RSC 2010				■								
RSC 2010 Conference							■					
Work on proposed extension (2)†								■	■	■	■	
Work on proposed extension (3)†												■

Abbreviations: APTS, Academy for PhD Training in Statistics; RSC, Research Students' Conference in Probability and Statistics.

† Refer to Chapter 5

Table 6.2: Diary of 2010/2011

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Work on proposed extension (3)†	■	■	■										
Work on proposed extension (4)†				■	■	■	■						
Prepare abstract for ISCB 2011					■								
Modelling with real-life data								■	■				
Thesis write-up										■	■	■	■
ISCB 2011 Conference											■		
Thesis submission													■

Abbreviations: ISCB, International Society for Clinical Biostatistics Conference.

† Refer to Chapter 5

Bibliography

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician*, 46(2):261–283.

A'Hern, R. P. (2001). Sample size tables for exact single-stage phase II designs. *Statistics in Medicine*, 20(6):859–866.

Berger, J. O. (1980). *Statistical decision theory: foundations, concepts, and methods*. Springer-Verlag, New York.

Brunier, H. C. and Whitehead, J. (1994). Sample sizes for phase II clinical trials derived from Bayesian decision theory. *Statistics in Medicine*, 13(23-24):2493–2502.

Bull, J. P. (1959). The historical development of clinical therapeutic trials. *Journal of Chronic Diseases*, 10(3):218–248.

Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 16(23):2701–2711.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press. Cambridge.

Crowley, J. and Ankerst, D. P., editors (2006). *Handbook of statistics in clinical oncology*. Chapman & Hall/CRC, Boca Raton, Second edition.

Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*, 13(17):1727–1736.

Fitzpatrick, S. (2005). *Clinical trial design*. Institute of Clinical Research, [S.l.].

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1998). *Fundamentals of clinical trials*. Springer-Verlag, New York, Third edition.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, 13(4):346–353.

Green, S., Benedetti, J., and Crowley, J. (2003). *Clinical trials in oncology.* Chapman & Hall/CRC, Boca Raton, Second edition.

Hackshaw, A. K. (2009). *A concise guide to clinical trials.* Wiley-Blackwell/BMJ Books, Chichester, UK.

International Conference on Harmonisation of Technical Requirements fro Registration of Pharmaceuticals for Human Use (ICH) (1998). Statistical principles for clinical trials E9.

Julious, S. A. and Swank, D. J. (2005). Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan. *Pharmaceutical Statistics*, 4(1):37–46.

Lee, S. J. and Zelen, M. (2000). Clinical trials and sample size considerations: another perspective. *Statistical Science*, 15(2):95–103.

Leung, D. H.-Y. and Wang, Y.-G. (2001). Optimal designs for evaluating a series of treatments. *Biometrics*, 57(1):168–171.

Lilienfeld, A. (1982). The Fielding H. Garrison Lecture: Ceteris paribus: the evolution of the clinical trial. *Bulletin of the History of Medicine*, 56(1):1–18.

Lindley, D. V. (1971). *Bayesian statistics: a review.* Regional conference series in Applied Mathematics, 2. Society for Industrial and Applied Mathematics, Philadelphia.

Machin, D., Campbell, M. J., Tan, S., and Tan, S. (2009). *Sample size tables for clinical studies.* Wiley-Blackwell, Chichester, UK, Third edition.

Machin, D., Simon, D., and Green, S. B. (2006). *Textbook of clinical trials.* J. Wiley & Sons, Chichester, West Sussex, England, Second edition.

Mariani, L. and Marubini, E. (1996). Design and analysis of phase II cancer trials: a review of statistical methods and guidelines for medical researchers. *International Statistical Review / Revue Internationale de Statistique*, 64(1):61–88.

Meinert, C. L. and Tonascia, S. (1986). *Clinical trials: design, conduct, and analysis.* Monographs in Epidemiology and Biostatistics, Volume 8. Oxford University Press, New York.

O'Hagan, A. and Stevens, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21(3):219–230.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46(1):33–48.

Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research*, 12(6):489–504.

Rice, J. A. (1995). *Mathematical statistics and data analysis.* Duxbury Press, Belmont, CA, Second edition.

Schoenfeld, D. (1980). Statistical considerations for pilot-studies. *International Journal of Radiation Oncology Biology Physics*, 6(3):371–374.

Senn, S. (1996). Some statistical issues in project prioritization in the pharmaceutical industry. *Statistics in Medicine*, 15(24):2689–2702.

Silvey, S. D. (1975). *Statistical inference.* Monographs on statistical subjects. Chapman and Hall, London.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1):1–10.

Simon, R., Wittes, R. E., and Ellenberg, S. S. (1985). Randomized phase II clinical trials. *Cancer Treatment Reports*, 69(12):1375–1381.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics*, 54(1):279–294.

Stallard, N. (2008). Phase II clinical trials. In Biswas, A., Datta, S., Fine, J. P., and Segal, M. R., editors, *Statistical advances in the biomedical sciences: clinical trials, epidemiology, survival analysis, and bioinformatics*, Wiley series in probability and statistics, chapter 2, pages 15–31. Wiley-Interscience, Hoboken, N.J.

Sylvester, R. J. (1988). A Bayesian approach to the design of phase II clinical trials. *Biometrics*, 44(3):823–836.

Tan, S.-B. and Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine*, 21(14):1991–2012.

Thall, P. F. and Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*, 50(2):337–349.

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A. T., Christian, M. C., and Gwyther, S. G. (2000). New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3):205–216.

Whitehead, J. (1985). Designing phase II studies in the context of a programme of clinical research. *Biometrics*, 41(2):373–383.

Whitehead, J. (1986). Sample sizes for phase II and phase III clinical trials: an integrated approach. *Statistics in Medicine*, 5(5):459–464.

Yao, T.-J., Begg, C. B., and Livingston, P. O. (1996). Optimal sample size for a series of pilot trials of new agents. *Biometrics*, 52(3):992–1001.