成蹊大学大学院　理工学研究科 博士

学位論文

ADAPTIVE DESIGN AND STATISTICAL INFERENCE IN CLINICAL TRIALS

2014 年 1 月

東郷 香苗

CONTENTS

1 INTRODUCTION 1
1.1 Backgrounds 1
1.2 Outline of each section 5
2 OPTIMAL TIMING FOR INTERIM ANALYSES IN CLINICAL TRIALS 8
2.1 INTRODUCTION 8
2.2 GENERAL FRAMEWORK 10
2.2.1 Sample size determination 10
2.2.2 Control of subject enrollment 11
2.3 METHODS 12
2.3.1 Average sample number for Case 1 12
2.3.2 Average sample number for Case 2 12
2.3.3 Subject enrollment model 14
2.3.4 Sample size adjustment 14
2.4 RESULTS 16
2.4.1 Results of numerical calculations in Case 1 17
2.4.2 Sample size based on the minimal clinically important effect size in Case 1 18
2.4.3 Results of numerical calculations in Case 2 19
2.4.4 Results of an interim analysis allowing sample size adjustment 20
2.5 DISCUSSION 21
3 SAMPLE SIZE RE-ESTIMATION FOR SURVIVAL DATA IN CLINICAL TRIALS WITH AN ADAPTIVE DESIGN 24
3.1 INTRODUCTION 24
3.2 GROUP-SEQUENTIAL TRIALS WITH SAMPLE SIZE RE-ESTIMATION 26
3.2.1 Assumptions 26
3.2.2 Methods to Preserve Type I Error Rate 27
3.3 NUMBER OF EVENTS AND SAMPLE SIZE REQUIRED IN THE SECOND STAGE 29
3.3.1 Target Number of Events and Sample Size in the Second Stage 29
3.3.2 Method to Estimate Hazards 31
3.4 SIMULATION 32
3.5 EXAMPLE 36
3.6 DISCUSSION 37

4 CLINICALLY IMPORTANT EFFECTS IN NEW DRUG DEVELOPMENT 39
4.1 INTRODUCTION 39
4.2 MCIC AND MCID 40
4.3 ROLES OF MCIC AND MCID IN CLINICAL TRIALS 42
4.4 APPROACHES TO MCIC AND MCID 43
4.4.1 DISTRIBUTION-BASED APPROACH 43
4.4.2 ANCHOR-BASED APPROACH 44
4.4.3 OPINION-BASED APPROACH 44
4.5 INTERPRETATION OF TRIAL RESULTS FOR CLINICAL IMPORTANCE 45
4.6 DISCUSSION AND CONCLUSION 47
5 GROUP COMPARISONS INVOLVING ZERO-INFLATED COUNT DATA IN CLINICAL TRIALS 48
5.1 Introduction 48
5.2 Two-part statistics and sample size 49
5.3 Comparison 52
5.3.1 Comparison of two-part statistics using Wilcoxon test, Wilcoxon test adjusted for ties, and *t*-test 52
5.3.2 Comparison with conventional tests 53
5.3.3 Comparison with the ZIP model 56
5.4 Example 59
5.5 Discussion 60
6 CONCLUSION 62
ACKNOWLEDGEMENT 65
REFERENCES 66

1 INTRODUCTION

1.1 Backgrounds

Clinical development of new drugs needs long time and huge costs. Before a new drug is approved by the regulatory authority for selling, efficacy and safety of the drug must be carefully evaluated as well as the quality of the drug because the approval of drugs greatly affects national health and also national finance. Drug development is ideally a logical, step-wise procedure in which information from small early studies is used to support and plan later larger, more definitive studies (ICH-E8, 1997) (Figure 1-1). In exploratory studies of early phases (phase 1 to 2), the statistical analyses such as modeling and visualization using collected data are frequently used. In confirmatory studies of late phases (phase 2 to 3; Figure 1-2), statistical testing for pre-defined hypotheses directly leads the conclusion of success or fails of the study. Therefore, drug development is one of the fields where statistical expertise is most required.

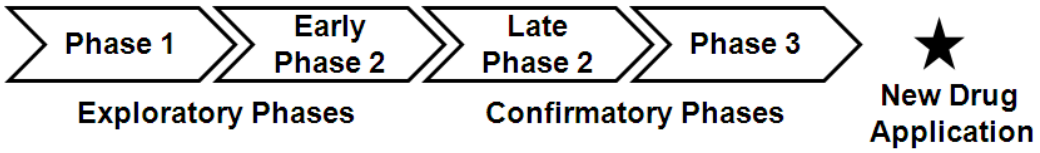


Figure 1-1. Schema of clinical development

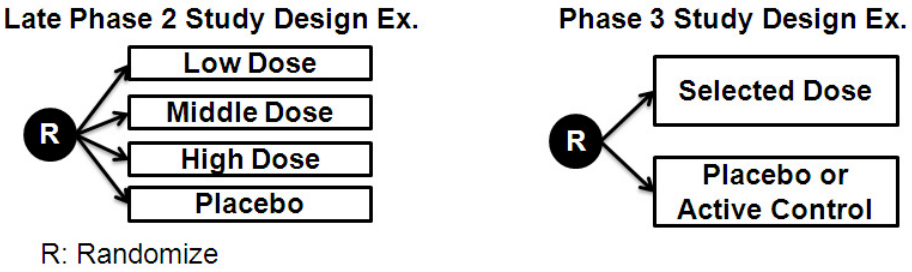


Figure 1-2. Examples of confirmatory study design

For the last decade, the scale of phase 3 studies have been getting large because the difference in the effect size between a new drug and the control drug, which should be demonstrated to be statistically and clinically significant in phase 3 studies, has been getting small. In addition, regulatory authorities request large safety data. The drug development is becoming increasingly expensive, and has a high clinical failure rate.

Pharmaceutical companies try to overcome these difficulties in various ways: one is adaptive design of clinical studies. Conventional design of clinical studies, especially confirmatory studies, is determined in details before the study starts, and does not change depending on the interim results in order to assure the validity and integrity of the study (Figure 1-3). In contrast, adaptive design allows us to modify the study design based on the interim results of the ongoing study by predefined decision rules (Figure 1-4). Adaptive design has the potential to speed up the process of drug development or can be used to allocate resources more efficiently without lowering scientific and regulatory standards (EMA, 2007). Gallo, et al. (2006) described adaptive design as follows:

"*By adaptive design we refer to a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial. The goal of adaptive designs is to learn from the accumulating data and to apply what is learned as quickly as possible. In such trials, changes are made "by design," and not on an ad hoc basis; therefore, adaptation is a design feature aimed to enhance the trial, not a remedy for inadequate planning.*"

There are many types of adaptive design: e.g., discontinuing treatment arms, seamless phase 2/3 design, and sample size re-estimation. It is a kind of traditional adaptive design to perform interim analyses for an early termination of clinical trials due to overwhelming drug effects or the futility of trials. These interim analyses have been used in oncology clinical trials for many years because cancer is life-threatening disease and most anti-cancer drugs have severe toxicity. These days, many phase 3 confirmatory trials in other therapeutic areas also employ the interim analyses for an early termination. They may also be planned for sample size re-estimation at interim analyses. The inducements to plan such interim analyses are a cost cutback and speedup of new drug developments as well as the ethical standpoint of avoiding exposure to study drugs beyond necessity.

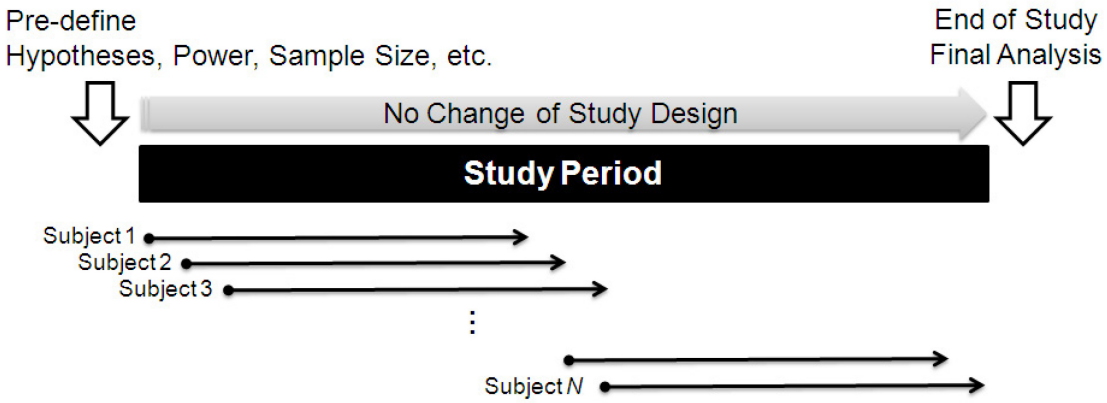


Figure 1-3. Conventional study design

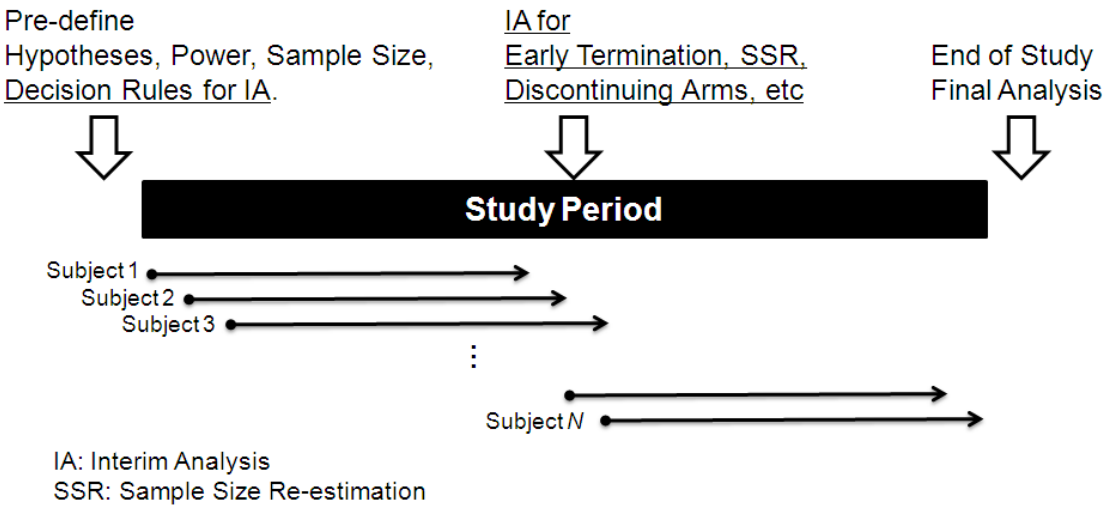


Figure 1-4. Adaptive design

The design used to re-estimate the sample size on the basis of the observed treatment effect have been controversial form various viewpoints such as complicated inferential decisions, the possibility of resulting in clinically meaningless differences, and efficiency (Shi, 2001; Shih, 2006; Taiatis, 2003). Nevertheless, the sample size re-estimation itself has attracted the attention of many investigators, since it is common that uncertainties remain in the critical assumptions about the effect size and extent of data variation during the design of a trial before its start. In the case of continuous and binary data, the methodology for sample size re-estimation has been intensively discussed for two decades. Proschan and Hunsberger (1995) proposed a conditional power based on the treatment effect in an interim analysis and proposed critical values

to preserve the type I error rate for the test of a two-sample mean. Cui, Hung and Wang (1999) proposed a test statistic for testing the two-sample means that preserve the type I error rate, using the ordinary critical values of fixed group sequential designs.

When the sample size re-estimation is implemented, the upper limit of the sample size should be set because huge clinical trials are infeasible and expose many people to study drugs which have perhaps no efficacy benefit. It is recommended to set the upper limit of the sample size based on the clinical importance (Shih, 2001; Hung *et al*., 2006). Futhermore, demonstrating the clinical importance of the drug effect should be a key element of efficacy evaluations in addition to simply showing a statistical significance. The clinical importance or clinical meaningfulness of the new investigational drug is defined and assessed generally by comparing drug effects with the Minimal Clinically Important Change from baseline (MCIC) or Minimal Clinically Important Difference between groups (MCID) of a primary endpoint. While the clinical importance of drug effect is often evaluated when interpreting results of clinical trials, it is also important to consider the clinical importance at the planning phase. However what is the widely accepted definition of the clinically important effect in the first place? There are various definitions, such as: changes in restoring normal levels of functions by the end of drug therapy, changes in significantly reducing patients' risk for various health problems, changes in a level of what is recognized and accepted as "improvement" by doctors and patients, and so on (Jacobson and Truax, 1991).

One of the purposes of this article is to deal with three subjects relative to adaptive design. First, the optimal time for interim analyses is examined. Any adaptive design needs interim analyses. The time of conducting an interim analysis affects the probability of the early termination and the number of subjects enrolled until the interim analysis. Second, we address the methodology of sample size re-estimation. Among many types of adaptive design, sample size re-estimation has attracted the attention of many investigators, since it is common that uncertainties remain in the critical assumptions about the effect size and extent of data variation in the planning stage. We address the methodology of sample size re-estimation for survival data which is still in developing phase, and propose an interim hazard ratio estimate that can be used to re-estimate the sample size under those circumstances. Third, clinically important effects are described. As mentioned above, clinically important effects are deeply related to sample size re-estimation. We provide the concept of MCIC and MCID and make clear the way to use them in new drug development through resolving the issues or uncertainties relative to MCIC and MCID.

Another purpose of this article is to present methods of statistical testing and sample size for a special distribution. Although adaptive design is a hot topic for efficient drug development, it is more essential for the efficient clinical trials to choose a statistical method appropriate to the distribution. Most statistical tests assume normal distribution for raw data or rank-transformed data. That is known as a robust approach even if the actual distribution is not quite fit to the normal distribution. However, special distributions like a bimodal distribution need different methods. For example, in clinical trials, outcomes of count data sometimes have excess zeros. These outcomes may be found in the number of symptoms (e.g., urinary incontinence episodes, gastrointestinal ulcers, hot flushes arising from menopausal disorder), the number of events (e.g., hospitalizations, heart attacks), and questionnaire scores. When the treatment difference between a test drug and a control is tested, either zero-inflation or the difference in the non-zero part is sometimes ignored. For example, a nonparametric test may be used after applying rank transformation to the data. This poses the problem that there can be many ties due to zero-inflation. In this case, the normal approximation is not accurate, whereas many nonparametric tests use the normal approximation for rank data. In addition, the power to detect the treatment difference could be decreased if there are many ties in the two treatment groups. Another example is that the treatment groups are compared only in the proportion of subjects with zero as a dichotomous response. This possibly wastes the treatment difference in the non-zero part by ignoring that. This response variable follows the binominal distribution in the zero part and the zero-truncated count distribution in the non-zero part. By applying two-part model, Lachenbruch (2001a) proposed a test statistic called the two-part statistic that combines the test statistics of the zero and non-zero parts. We provide methods for finding the sample size and power for the two-part statistic. Furthermore, the power of the two-part statistic is examined compared with the conventional methods and the zero-inflated Poisson model.

1.2 Outline of each section

In Chapter 2, the optimal time for interim analyses is addressed. In clinical trials, interim analyses are often performed before the completion of the trial. The intention is to possibly terminate the trial early or adjust the sample size. The time of conducting an interim analysis affects the probability of the early termination and the number of subjects enrolled until the interim analysis. This influences the expected total number of

subjects. In this study, we examine the optimal time for conducting interim analyses with a view to minimizing the expected total sample size. It is found that regardless of the effect size, the optimal time of one interim analysis for the early termination is approximately two-thirds of the planned observations for the O'Brien–Fleming type of spending function and approximately half of the planned observations for the Pocock type when the subject enrollment is halted for the interim analysis. When the subject enrollment is continuous throughout the trial, the optimal time for the interim analysis varies according to the follow-up duration. We also consider the time for one interim analysis including the sample size adjustment in terms of minimizing the expected total sample size.

Chapter 3 addresses the methodology of sample size re-estimation for survival data which is still in developing phase. In clinical trials with survival data, investigators may wish to re-estimate the sample size based on the observed effect size while the trial is ongoing. Besides the inflation of the type I error rate due to sample size re-estimation, the method for calculating the sample size in an interim analysis should be carefully considered because the data in each stage are mutually dependent in trials with survival data. Although the interim hazard estimate is commonly used to re-estimate the sample size, the estimate can sometimes be considerably higher or lower than the hypothesized hazard by chance. We propose an interim hazard ratio estimate that can be used to re-estimate the sample size under those circumstances. The proposed method was demonstrated through a simulation study and an actual clinical trial as an example. The effect of the shape parameter for the Weibull survival distribution on the sample size re-estimation is presented.

Chapter 4 describes clinically important effects. As mentioned above, clinically important effects are deeply related to sample size re-estimation. In new drug development, demonstrating a clinically important effect of the new drug is a key element of efficacy evaluations instead of simply showing a statistical significance. However, approaches to demonstrate clinically important effects are unclear and not well recognized among many investigators and sponsors. The Minimal Clinically Important Change from baseline (MCIC) and Minimal Clinically Important Difference between groups (MCID) are used to assess the clinically important effect. We state the roles of MCIC and MCID in each phase of new drug development and common approaches to establishing MCIC and MCID. Furthermore, we provide some common approaches on how to practically compare the clinical trial results with the MCIC and MCID and interpret the clinical importance from applications in new drug development.

We also suggest incorporating the clinical importance at the planning phase of trials.

Chapter 5 presents statistical testing and sample size for zero-inflated count data in clinical trials. In clinical trials, outcomes of count data sometimes have excess zeros. When a test drug is compared to a control, zero-inflated data may be ignored or interest is taken only in the proportion of zero counts. By applying the two-part model, Lachenbruch (2001a) suggested a test statistic called the two-part statistic that combines the test statistics of the zero part and the non-zero part. The test for the zero part is the chi-square test. The test for the non-zero part may be a Wilcoxon test, a t-test, etc. This article proposes methods for calculating the sample size and power for the two-part statistic with zero-inflated Poisson data. We developed the methods of sample size and power for the two-part statistic using the Wilcoxon test adjusted for ties. The relationship between the non-zero part and zero-truncated Poisson distribution is also described. Furthermore, we examine the power of the two-part statistic, conventional methods, and the zero-inflated Poisson model.

Chapter 2 is based on Togo and Iwasaki (2013) with some revisions. Chapter 3 contains the revised article of Togo and Iwasaki (2011). Chapter 4 is based on Togo, Matsuoka, et al. (2013) and Chapter 5 is based on Togo and Iwasaki (2013) with some revisions.

2 OPTIMAL TIMING FOR INTERIM ANALYSES IN CLINICAL TRIALS

2.1 INTRODUCTION

Interim analyses are often planned with the intention of an early termination of clinical trials because of various reasons, such as overwhelming drug effects or the futility of trials. They may also be planned for adjusting the sample size. The inducements to plan interim analyses are a cost cutback and speedup of new drug developments as well as the ethical standpoint of avoiding exposure to study drugs beyond necessity. A simple approach for the timing of interim analyses is to plan interim analyses at equal intervals such as half of the planned observations for one interim analysis, and one-third and two-thirds of the planned observations for two interim analyses. It is easy to find clinical trials conducting interim analyses at equal intervals. However, there is little probability to terminate the clinical trial early if the interim analysis at half of the planned observations has too little information. Such interim analyses waste type I error spending and aren't worth operational and cost burdens caused by the interim analyses. Other timing might be chosen if criteria appropriate for choosing the timing are defined. In terms of reducing costs in confirmatory trials, the optimal time for interim analyses can be chosen so that it minimizes the expected sample size which is frequently called Average Sample Number (ASN) (Colton and McPherson, 1976). This approach may not work for early phase trials because the trial objectives include predicting the effect size of the treatment for the following trials, whereas confirmatory trials can focus on demonstrating the superiority or non-inferiority. The interim analysis timing in the proof of concept trials is examined by Gould (2005).

A common index for the time of the interim analysis is the information time (or information fraction) based on test statistics (Proschan *et al*., 2006). When the endpoint is a normally distributed outcome, the information time is the proportion of the number of subjects observed until the interim analysis to the pre-planned sample size. On the other hand, the ASN has relevance not only to the information time but also to the number of subjects enrolled in the clinical trial until the interim analysis. There are two cases of controlling the subject enrollment. The first case is when the subject enrollment halts as soon as the target number of subjects for an interim analysis is reached. The second case is when the enrollment is continuous throughout the trial. In the second case, not all subjects enrolled until the interim analysis are used for the interim analysis,

but only subjects who complete the follow-up period for the primary endpoint are used. The enrollment of subjects who have been enrolled but are not used for the interim analysis would be fruitless if the clinical trial is terminated early due to futility based on the results of the interim analysis. However, in late phase clinical trials, the second case is common because the speed of development of the new drug is critical. The choice of the first or second case depends on the development strategy and circumstances of the disease. When the second case is chosen, the estimation of the ASN needs to expect the number of subjects enrolled until the interim analysis, which is determined by the follow-up duration and the enrollment rate. Therefore, those two factors should be considered in the planning stage.

Another point to be considered is that the ASN depends on the approach to the pre-planned sample size. The sample size is frequently planned in order to have the desired power to detect the anticipated effect size in the treatment difference. However, many textbooks state that the sample size should have the power to detect the minimal clinically important effect because sponsors and investigators tend to overestimate the treatment effect (e.g., Chow and Liu, 2004; Jennison and Turnbull, 2006). Let us assume that the minimal clinically important effect size is Δ and the anticipated effect size is L times greater than Δ ($L \geq 1$). If the sample size is determined on the basis of Δ, despite anticipating a large L, it is inefficient in terms of demonstrating the statistical significance in the treatment difference (Golub, 2006). In this case, a group sequential design including an early termination is effective. It can be said that the sample size adjustment is the opposite approach; the sample size is pre-planned on the basis of $L\Delta$, and then the sample size can be adjusted up to that on the basis of Δ (Hung et al., 2006).

This article addresses the optimal time for interim analyses in terms of minimizing the ASN in confirmatory trials when the interim analysis allows an early termination or a sample size adjustment. Furthermore, we discuss the duration of the follow-up from a viewpoint of cost-effectiveness of an interim analysis. In clinical trials with a long duration of the follow-up, the ASN in a group sequential design may be little less than the sample size required in a single stage design. Section 2.2 presents the general framework of this study. Two cases of controlling subject enrollment are introduced. Section 2.3 focuses on the methodology of calculating the ASN for each case of subject enrollment. The methods for the ASN adopt the sample size based on $L\Delta$. In addition, subject enrollment models are presented in order to estimate the ASN for the case of continuous enrollment. Section 2.4 shows the results of the numerical calculations of the ASN and the optimal time for the interim analysis. The impact of a long follow-up

duration on the ASN is demonstrated. Some discussions and concluding remarks are given in Section 2.5.

2.2 GENERAL FRAMEWORK

We consider a clinical trial to compare a test drug with a control. Assume that outcomes in the test drug and the control follow $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with known variance, respectively. The null and alternative hypotheses to be tested are

$$H_0: \theta = 0 \quad \text{vs.} \quad H_1: \theta > 0$$

where θ is the standardized treatment difference of $\mu_1 - \mu_2$.

Assume that K analyses (i.e., $K - 1$ interim analyses) are planned. The decision rule for an interim analysis is as follows:

A) If $Z_k > b_k$, then reject H_0 (early termination for success),

B) If $Z_k \le c_k$, then accept H_0 (early termination for futility), and

C) If $c_k < Z_k \le b_k$, then continue the trial and enroll n_{k+1} subjects in k+1th stage. When $k = K$, stop the trial and accept H_0.

Here, Z_k is a test statistic ~ N(0, 1) under the null hypothesis and b_k and c_k are the upper and lower boundaries of the test at the k^{th} interim analysis for k=1, …, K. Let N and n_k denote the pre-planned maximum number of subjects and the number of subjects enrolled at the k^{th} stage, respectively. The information time for the k^{th} interim analysis t_k is $\Sigma n_k/N$.

2.2.1 Sample size determination

The pre-planned sample size N is determined so that the sample size has the desired power of $1 - \beta$. Under the alternative hypothesis of $\theta = \Delta$, the power is given by

$$pr\{Z_1 > b_1 | \theta = \Delta\} + \cdots + pr\{c_1 < Z_1 \le b_1; \cdots; c_{K-1} < Z_{K-1} \le b_{K-1}; Z_K > b_K | \theta = \Delta\}.$$

When K=2, the power is given by

$$\begin{aligned} &\int_{b_1 - z_1}^{\infty} \phi(x)dx + \int_{c_1 - z_1}^{b_1 - z_1} \int_{b_2 - z_2}^{\infty} \phi_2(x,y)dydx \\ &= 1 - \Phi_2(b_1 - z_1, b_2 - z_2) + \Phi_2(c_1 - z_1, b_2 - z_2) - \Phi(c_1 - z_1) \end{aligned},$$

where $z_1 = \Delta\sqrt{t_1 N/4}$ and $z_2 = \Delta\sqrt{N/4}$ under the alternative hypothesis, and

$\rho(z_1, z_2) = \sqrt{t_1}$. Here, ϕ and Φ are univariate standard normal density and distribution functions, respectively; and ϕ_2 and Φ_2 are bivariate standard normal density and distribution functions with coefficient correlation ρ, respectively. Calculation of the power to achieve the target power of $1 - \beta$ requires numerical integration.

2.2.2 Control of subject enrollment

Consider two cases of controlling the subject enrollment at interim analyses (Figure 5-1). Case 1 is when the enrollment halts after the pre-planned number of subjects Σn_k for the k^{th} interim analysis are enrolled, and the enrollment is resumed after the interim analysis. Case 2 is when the enrollment does not halt even after the number of subjects enrolled reaches Σn_k. In this case, we assume that the endpoint for the hypothesis test is assessed at a follow-up period of F. Let n_{k1} denote the number of subjects enrolled during F after the Σn_{k-1} subjects are enrolled and $n_{k2} = n_k - n_{k1}$.

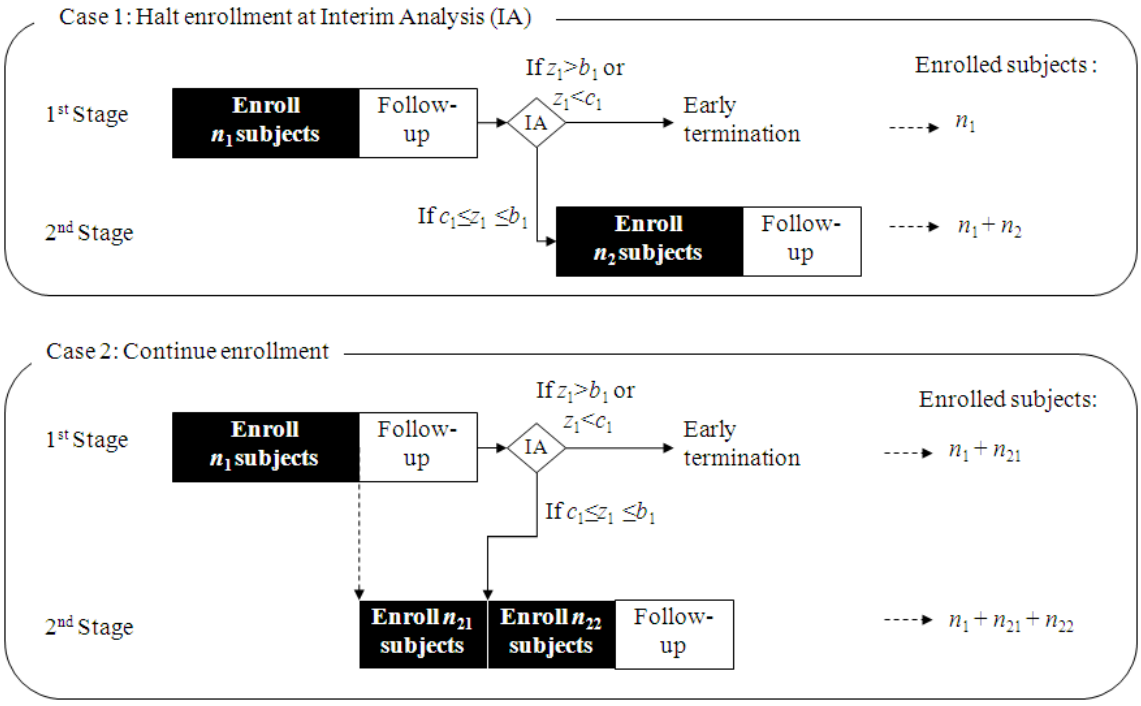


Figure 2-1. Schema of the supposed trial when $K = 2$.

2.3 METHODS

2.3.1 Average sample number for Case 1

In addition to equations of the ASN generalized to any K, we present equations for $K = 2$ because it is simple to understand and one interim analysis is most common. The ASN for $K - 1$ interim analyses is calculated as

$$\mathrm{ASN} = n_1 + \sum_{k=2}^{K} n_k \Pr\left(k^{th}\ stage\ required \middle| \theta = \Delta\right). \tag{2-1}$$

When $K = 2$,

$$\begin{aligned}\mathrm{ASN} &= n_1 + n_2 \Pr\left(c_1 < Z_1 \le b_1 \middle| \theta = \Delta\right) \\ &= N\left\{t_1 + (1 - t_1)\int_{c_1 - z_1}^{b_1 - z_1} \phi(x)dx\right\}\end{aligned}.$$

As mentioned in Section 2.1, the sample size is sometimes planned to achieve the power to detect the minimal clinically important effect Δ instead of the anticipated effect that is L times greater than Δ ($L \ge 1$). If the sample size is determined on the basis of Δ, the ASN under the assumption of $L\Delta$ is

$$\mathrm{ASN} = n_1 + \sum_{k=2}^{K} n_k \Pr\left(k^{th}\ stage\ required \middle| \theta = L\Delta\right). \tag{2-2}$$

2.3.2 Average sample number for Case 2

Let R denote an enrollment period and m denote a calendar time. The information time of an interim analysis, t, is described as

$$t = \begin{cases} 0 & (m < F) \\ g(m - F) & (F \le m \le F + R) \\ 1 & (F + R < m) \end{cases}.$$

These assume no withdrawal of subjects during the follow-up. Here, $g(m)$ is a function of the proportion of the number of enrolled subjects at m to N. Actual examples of $g(m)$ are presented in the next section. Figure 2-2 shows a subject enrollment model and a model of t.

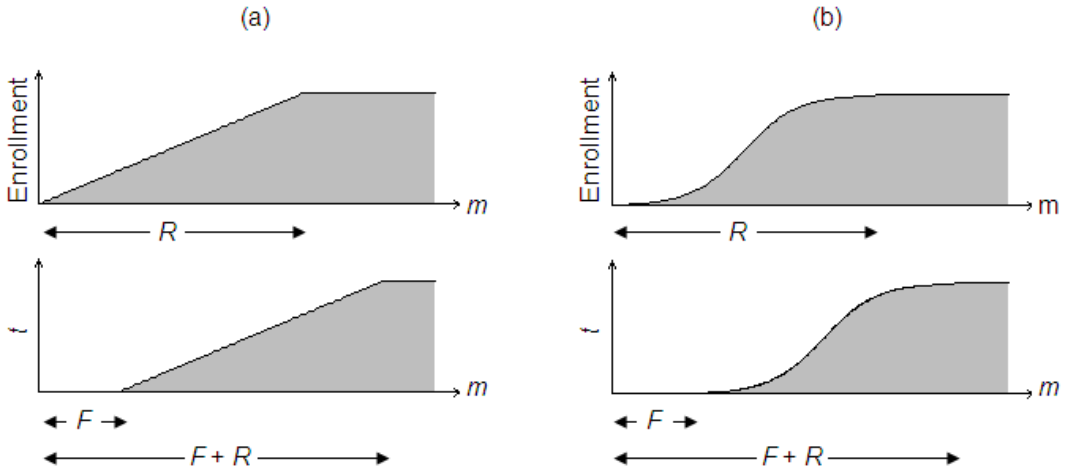


Figure 2-2. Subject enrollment (the upper figure) and information time (the lower figure) when *K* = 2: (a) subject enrollment is constant accrual; (b) subject enrollment is S-shaped.

When $K = 2$, the number of enrolled subjects until the interim analysis is $n_1 + n_{21}$. The ASN is calculated as

$$\mathrm{ASN} = n_1 + n_{21} + n_{22}\Pr\left(2^{nd}\ stage\ required\right). \tag{2-3}$$

If the interim analysis is conducted at $t_1 = n_1/N$ or $m = g^{-1}(t_1) + F$, the proportion of the enrolled subjects is $g(g^{-1}(t_1) + F)$. Therefore, n_{21} and equation (2-3) are written as

$$n_{21} = N \cdot g\left(g^{-1}(t_1) + F\right) - n_1,$$

and

$$\mathrm{ASN} = N\left\{g\left(g^{-1}(t_1) + F\right) + \left(1 - g\left(g^{-1}(t_1) + F\right)\right)\int_{c_1 - z_1}^{b_1 - z_1}\phi(x)dx\right\}.$$

When $K > 2$, n_{k1} and the ASN are expressed as

$$n_{k1} = N \cdot g\left(g^{-1}(t_{k-1}) + F\right) - \sum_{i=1}^{k-1} n_i,$$

and

$$\mathrm{ASN} = n_1 + n_{21} + \sum_{k=2}^{K-1}\left\{\left(n_{k2} + n_{(k+1)1}\right)\Pr\left(k^{th}\ stage\ required\right)\right\} + n_{K2}\Pr\left(K^{th}\ stage\ required\right),$$

where $n_{k2} = n_k - n_{k1}$.

The interim analysis should be conducted at least before the completion of patient enrollment in terms of reducing the number of enrolled subjects. That is, the information time t_1 should be chosen to satisfy $g(g^{-1}(t_1) + F) < 1$, that is

$$t_1 < \cdots < t_{K-1} < t_K < g(R - F). \tag{2-4}$$

2.3.3 Subject enrollment model

We consider three enrollment models to describe the cumulative number of subjects. The first model is when subjects are enrolled at a constant rate. The proportion of the number of subjects enrolled at m to the sample size N is written as a function as follows:

$$g(m) = \begin{cases} m/R & (0 \le m \le R) \\ 1 & (R \le m) \end{cases}. \tag{2-5}$$

In practice, not all investigational sites start the enrollment all together in large clinical trials. In this case, the enrollment model is approximately S-shaped and can be modeled using a sine curve as

$$g(m) = \begin{cases} \dfrac{\sin(\pi m/R - \pi/2)+1}{2} & (0 \le m \le R) \\ 1 & (R \le m) \end{cases}. \tag{2-6}$$

A logistic curve is not appropriate because $g(m) \neq 0$ for $m = 0$ and $g(m) \neq 1$ for $m = 1$.

Another familiar enrollment model is profiled with the truncated exponential distribution (Lachin and Foulkes, 1986), and $g(m)$ is

$$g(m) = \int_0^m \frac{\gamma e^{-\gamma y}}{1 - e^{-\gamma R}} dy \quad (\gamma \neq 0).$$

For $\gamma>0$, the enrollment rate is convex, whereas for $\gamma < 0$, the enrollment rate is concave.

2.3.4 Sample size adjustment

Suppose that the ASN is calculated at the planning stage of a clinical trial in order to find the optimal timing of an interim analysis as well as Section 2.3.1 and 2.3.2; however, the interim analysis includes an adjustment of the sample size. We consider $K = 2$ because two or more adjustments of the sample size are not practical. If an interim analysis allows a sample size adjustment, the type I error rate could inflate. One of several approaches to preserve the type I error rate is to adjust a test statistic at the final analysis (Cui et al., 1999).

The decision rule dealt with here is as follows:

A) If $Z_1 > b_1$, then reject H_0 (early termination for success),

B) If $Z_1 \le c_1$, then accept H_0 (early termination for futility), and

C) If $c_1 < Z_1 \le b_1$, then re-estimate n_2 and enroll re-estimated n_2^* subjects in the second stage.

The decision rules to perform the sample size re-estimation are various. For example, the sample size is re-estimated if the conditional power is greater than a certain value in addition to the above rule of (C) $c_1 < Z_1 \le b_1$ (Chen et al., 2004). The method proposed in this section can be easily applied to other criteria.

The sample size $N^* = n_1 + n_2^*$ can be re-estimated so that the conditional power based on the interim estimate of the effect size θ becomes greater than $1 - \beta$ (Proschan and Hunsberger, 1995). Because the interim estimate of θ is unknown at the planning stage, the ASN is calculated using an anticipated effect size Δ^* which differs from the anticipated effect size for the pre-planned sample size N. For example, Δ^* is the smallest treatment difference expected as the worst scenario, or the minimum clinically meaningful treatment difference. The ASN is given by

$$\begin{aligned}\mathrm{ASN} &= n_1 + n_{21} + E\left[n_2^* \middle| c_1 < Z_1 \le b_1\right]\Pr\left(c_1 < Z_1 \le b_1\right) \\ &= n_1 + n_{21} + \left\{\sum_{\nu=N_L-n_1}^{N_U-n_1}(\nu - n_{21})\Pr\left(n_2^* = \nu \middle| c_1 < Z_1 \le b_1\right)\right\}\Pr\left(c_1 < Z_1 \le b_1\right), \quad (2\text{-}7) \\ &= n_1 + n_{21} + \sum_{\nu=N_L-n_1}^{N_U-n_1}(\nu - n_{21})\Pr\left(n_2^* = \nu\right)\end{aligned}$$

where N_L and N_U are the upper and the lower limits of N^*, respectively. The upper bound is determined so that it is at least less than the realistic size in view of the budget and feasibility or the size needed to detect a minimum clinically meaningful difference (Shih, 2001; Hung et al., 2006). The lower bounds should be N if the sample size adjustment does not allow the reduction from the pre-planned sample size. In equation (2-7), n_2^* is a random variable with the probability

$$\Pr\left(n_2^* = \nu\right) = \begin{cases} \int_{c_1}^{A(\nu)} \Phi\left(z \middle| \theta = \Delta^*\right) dz & \left(\nu = N_U - n_1\right) \\ \int_{A(\nu+1)}^{A(\nu)} \Phi\left(z \middle| \theta = \Delta^*\right) dz & \left(N_L - n_1 < \nu < N_U - n_1\right) \\ \int_{A(\nu+1)}^{b_1} \Phi\left(z \middle| \theta = \Delta^*\right) dz & \left(\nu = N_L - n_1\right) \end{cases}, \quad (2\text{-}8)$$

$$\text{where} \quad A\left(n_2^*\right) = \left\{b_2 - \Phi^{-1}(\beta)\sqrt{\frac{n_2^*}{n_1 + n_2^*}}\right\}\sqrt{\frac{n_1}{n_1 + n_2^*}}.$$

The details of equation (2-8) are following. The conditional power (Proschan and Hunsberger, 1995) at an interim analysis is given by

$$1-\Phi\left(\frac{b_2\sqrt{n_1+n_2^*}-z_1\sqrt{n_1}-n_2^*\theta/2}{\sqrt{n_2^*}}\right)\geq 1-\beta\,.$$

When the treatment difference observed in the interim analysis Δ^* is used instead of the unknown θ, the conditional power is

$$1-\Phi\left(\frac{b_2\sqrt{n_1+n_2^*}-z_1\sqrt{n_1}-n_2^*\Delta^*/2}{\sqrt{n_2^*}}\right)\geq 1-\beta\,.$$

When the interim test statistic $z_1=\Delta^*\sqrt{n_1/4}$, then $\Delta^*=2z_1/\sqrt{n_1}$. The above inequality can be expressed as

$$\frac{b_2\sqrt{n_1+n_2^*}-z_1\sqrt{n_1}-z_1n_2^*/\sqrt{n_1}}{\sqrt{n_2^*}}\leq\Phi^{-1}(\beta)$$

$$\begin{aligned}z_1&\geq-\left\{\Phi^{-1}(\beta)-b_2\sqrt{\frac{n_1+n_2^*}{n_2^*}}\right\}\left(\sqrt{\frac{n_1}{n_2^*}}+\sqrt{\frac{n_2^*}{n_1}}\right)^{-1}\\&\geq\left\{b_2-\Phi^{-1}(\beta)\sqrt{\frac{n_2^*}{n_1+n_2^*}}\right\}\sqrt{\frac{n_1}{n_1+n_2^*}}\end{aligned}\,.$$

Therefore, the z_1 range possible for $n_2{}^*$ is

$$\left\{b_2-\Phi^{-1}(\beta)\sqrt{\frac{n_2^*}{n_1+n_2^*}}\right\}\sqrt{\frac{n_1}{n_1+n_2^*}}\leq z_1<\left\{b_2-\Phi^{-1}(\beta)\sqrt{\frac{n_2^*-1}{n_1+n_2^*-1}}\right\}\sqrt{\frac{n_1}{n_1+n_2^*-1}}\,.$$

2.4 RESULTS

We performed a numerical calculation of the ASN in order to find t_1 that would minimize the ASN, when t_1 ranged from 0.1 to 0.9 and the effect sizes ranged from 0.1 to 0.9. As a measure of the efficiency of the group sequential design, ASN/N_0 was calculated where N_0 was the sample size for a single-stage design without interim analyses. The overall significance level α was 0.025 (one-sided), and the target power 1 - β was 0.9. The O'Brien-Fleming (O-F) type and Pocock type of the spending functions (Lan and DeMets, 1983) were employed to control the type 1 error rate. Hereafter, the optimal t_1 will be the t_1 that minimizes the ASN.

2.4.1 Results of numerical calculations in Case 1

The ASN in equation (2-1) was calculated when $K = 2$. The numerical calculations show that the optimal t_1 was almost constant from the effect size of 0.1 to 0.9 (Table 2-1 and Figure 2-3). This implies that, for minimizing the ASN, it is best to conduct an interim analysis when t_1 is approximately 2/3 and 1/2 for the O-F type and the Pocock type, respectively.

Table 2-1. Summary of the optimal t_1 when effect sizes range from 0.1 to 0.9

		$K = 2$		$K = 3$	
		O-F	Pocock	O-F	Pocock
t_1	Median	0.66	0.49	0.55	0.35
	(Range)	(0.65–0.67)	(0.45–0.49)	(0.54–0.56)	(0.33–0.39)
t_2	Median	–	–	0.79	0.77
	(Range)	–	–	(0.74–0.80)	(0.74–0.90)
ASN/N_0	Median	82%	78%	77%	70%
	(Range)	(82–83%)	(78–78%)	(77–77%)	(70–70%)

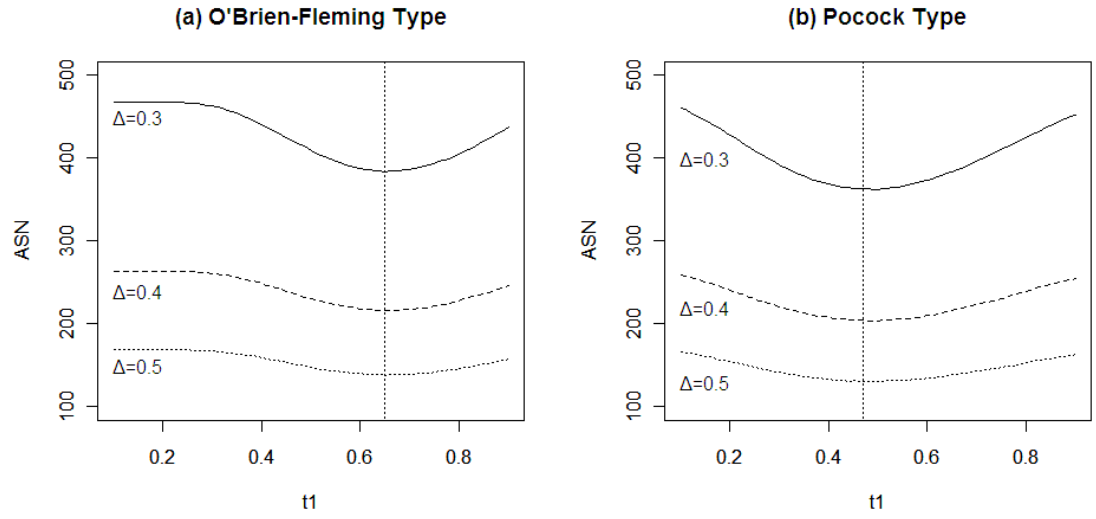


Figure 2-3. ASN versus t_1 when $K = 2$ and effect size (ES) ranges from 0.3 to 0.5.

For $K = 3$, the ASNs were calculated in the pairs of t_1 ($0.1 \leq t_1 \leq 0.8$) and t_2 ($t_1 + 0.1 \leq t_2 \leq 0.9$). The optimal t_1 and t_2 were summarized in Table 2-1. Although the

optimal t_2 varied widely for the effect size ranging from 0.1 to 0.9, t_2 had little effect on the ASN (Figure 2-4). Therefore, the ASN was almost minimal at the median t_1 and t_2 in Table 2-1 regardless of the effect size.

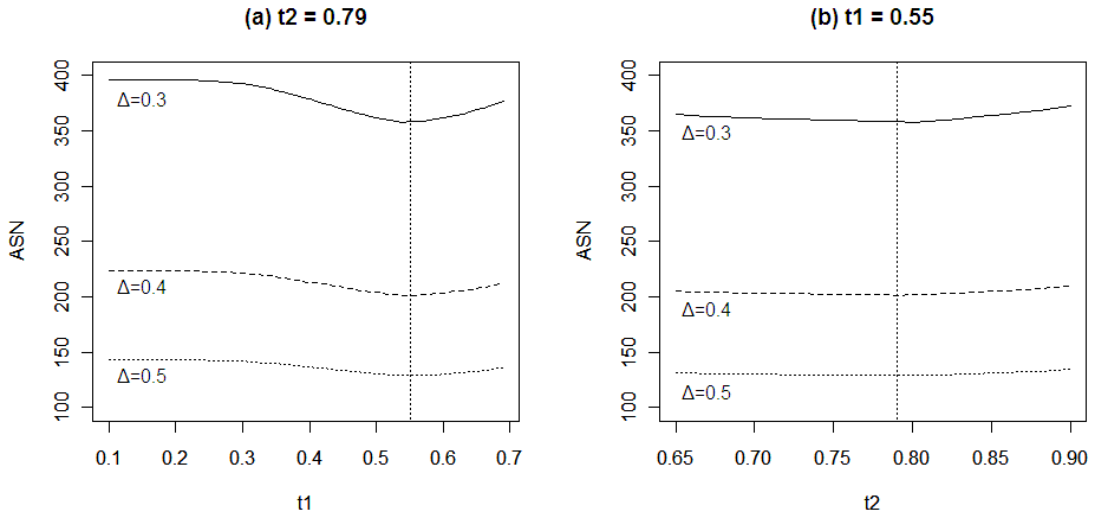


Figure 2-4. ASN versus t_1 and t_2 with the spending function of O'Brien–Fleming type when K = 3 and effect size (ES) = 0.3.

2.4.2 Sample size based on the minimal clinically important effect size in Case 1

We calculated the ASN in equation (2-2) when the sample size was pre-planned based on the minimal clinically important effect size Δ instead of the anticipated effect size $L\Delta$. Figure 2-5 shows that the median of the optimal t_1 and the minimal ASN/N_0 were getting smaller as L was larger. When L was 1.5, the optimal t_1 was approximately 0.1 less than that for L of 1 shown in Table 2-1.

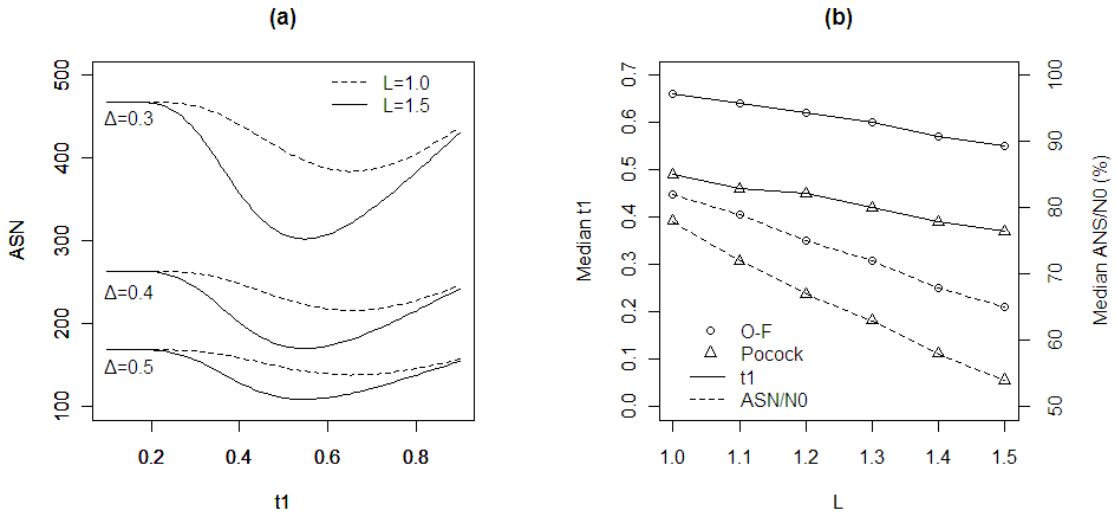


Figure 2-5. ASN and t_1 when the sample size is calculated based on Δ. The anticipated effect size is $L\Delta$ and $K = 2$: (a) t_1 versus ASN with the spending function of O'Brien–Fleming type; (b) L versus median of the optimal t_1 and median ASN/N_0 for the effect size (ES) in the 0.1–0.9 range.

2.4.3 Results of numerical calculations in Case 2

For Case 2, we assumed R was 12 months and F ranged from 1 to 4 months (i.e., $F/R = 0.08$ to 0.33). The ASN was calculated when the effect size ranged from 0.1 to 0.9, and the subject enrollment model $g(m)$ was based on a constant rate in equation (2-5). Figure 0-6(a) shows that the optimal t_1 decreased as F increased when using the O'Brien-Fleming type. In the Pocock type, the optimal t_1 was approximately 15% less than that for the O'Brien-Fleming type for every F and L.

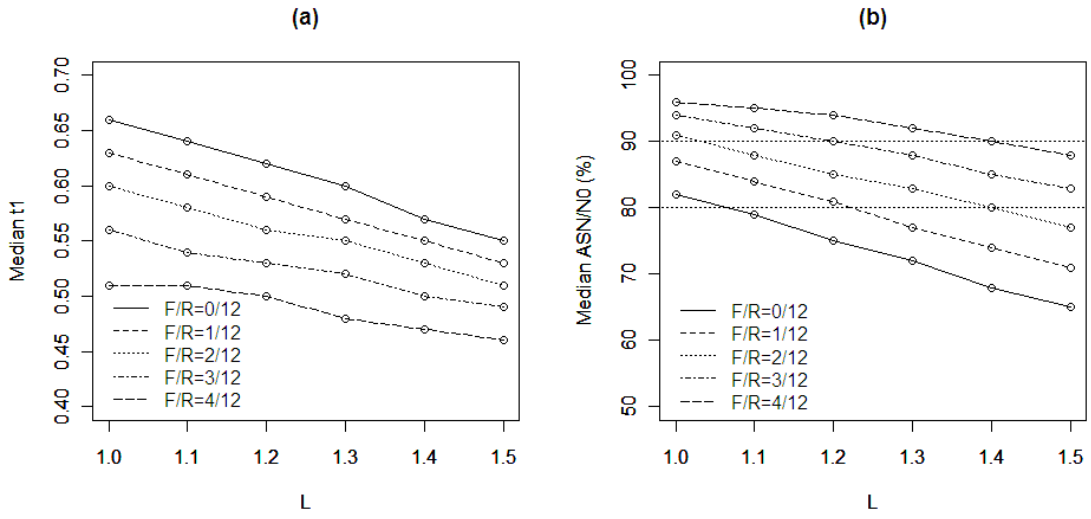


Figure 2-6. Median t_1 and ASN/N_0 at the minimal ASN through the effect size of 0.1-0.9 when the subject enrollment distribution is constant and K=2

Figure 2-6(b) shows the median ASN/N_0 for the effect sizes ranging from 0.1 to 0.9. The medians at L = 1 were over 80% for any F/R and over 90% for $F/R \geq 2/12$. Because interim analyses are cost and operational burdens, the ASN/N_0 should be small enough to motivate investigators and sponsors to perform interim analyses unless the primary purpose of the interim analysis is safety monitoring. For example, when 90% or less ASN/N_0 is the target reduction for deciding whether or not an interim analysis is to be planned, an interim analysis should not be planned in clinical trials of $F/R \geq 2/12$ at L = 1 based on t the numerical calculations (Figure 2-6(b)). However, ASN/N_0 decreased as L increased. When the sample size was calculated based on the minimal clinically important effect size Δ instead of the anticipated effect size $L\Delta$, the ASN/N_0 values in most cases were less than 90%. Nevertheless, it may be ineffective to conduct an interim analysis when $F/R \geq 4/12$.

2.4.4 Results of an interim analysis allowing sample size adjustment

We calculated ASN for t_1 ranging from 0.1 to 0.9 when the sample size was adjusted at an interim analysis. The pre-planned sample size N was calculated based on the anticipated effect size $L\Delta$ whereas the minimal clinically important effect size was Δ. It is the opposite of the interim analysis for the early termination in which the sample

size is pre-planned based on Δ, and the study is terminated if the $L\Delta$ is greater than Δ. Let us assume that the true effect size was $L^*\Delta$; i.e., $\Delta^* = L^*\Delta$ in equation (2-7). The sample size was allowed to increase with an upper limit of the sample size based on Δ.

Figure 2-7(a) shows the ASN and t_1 for $\Delta = 0.3$, $L = 1.5$, and Case 1 (i.e., $F = 0$). The curves for $L^* = 1.1$ to 1.5 were similar in shape. Regardless of L^*, the optimal t_1 was approximately 0.75 for the O'Brien-Fleming type and approximately 0.60 for the Pocock type for the effect sizes ranging from 0.1 to 0.9.

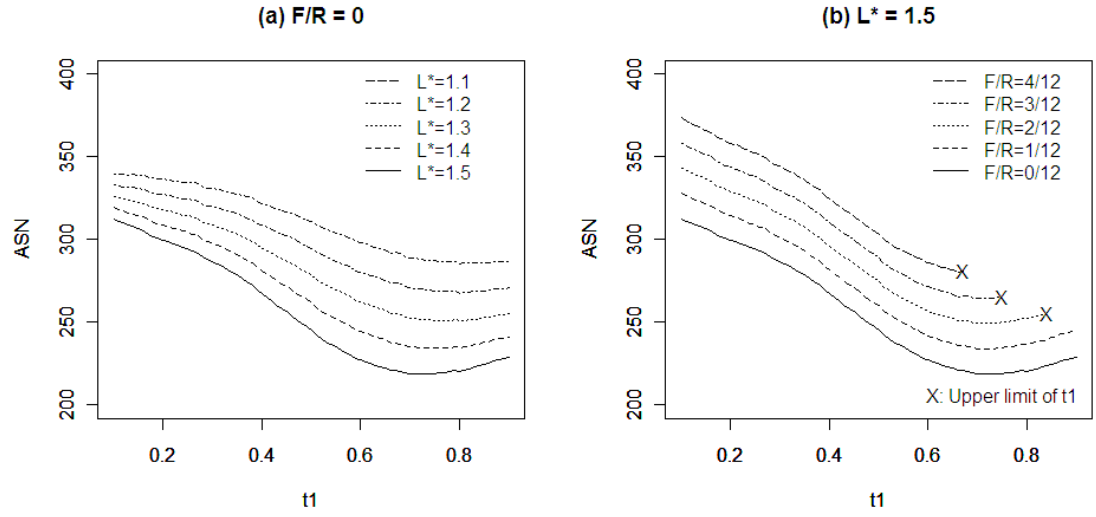


Figure 2-7. Comparison between S-shaped enrollment model and the enrollment model of a constant rate when effect size=0.3, K=2 and L=1

When F/R was from 1/12 to 4/12 in Case 2, all the ASN curves had the same shape (Figure 2-7(b)); this was also obvious from equation (2-7). Therefore, the optimal t_1 at any F/R should have been the same as Case 1. However, when the optimal t_1 in Case 1 was greater than the upper limit of t_1 in inequality (2-4), the optimal t_1 in Case 2 came down to the upper limit.

2.5 DISCUSSION

We demonstrated the methods used to find the optimal time for interim analyses in order to minimize the ASN. When an interim analysis was performed in Case 1, the optimal time was approximately $t_1 = 2/3$ for the O'Brien-Fleming type and approximately $t_1 = 1/2$ for the Pocock type, regardless of the effect size. These results

based on the ASN were consistent with the common impression that the optimal time for the O'Brien-Fleming type would be later than that for the Pocock type because of spending less type I error rate at the interim analysis with the O'Brien-Fleming type. When two interim analyses were performed in Case 1, the time of the second interim analysis had little effect on the ASN. The determination of the appropriate number of interim analyses is beyond the scope of this article; however, planning more than one interim analysis requires careful consideration unless the purpose of the interim analyses is safety monitoring.

When the sample size is planned based on the clinically meaningful effects, the optimal time of the interim analysis is earlier than when the sample size is planned based on the anticipated effect. The group sequential design is more efficient compared with a fixed design when the anticipated effect is much greater than the clinically meaningful effect.

We showed that the optimal time for the interim analysis depended on the follow-up duration in Case 2. The interim analysis did not considerably reduce the ASN when the follow-up duration was longer than one-third of the enrollment duration. Section 2.4 presented the results in the enrollment model using a constant rate in equation (2-5) alone. Regarding the S-shaped enrollment model in equation (2-6), the optimal t_1 is almost the same as those for the enrollment model of a constant rate; however the minimal ASN is smaller because of the slow enrollment rate at the beginning. It is important to take the enrollment duration and the subject enrollment model into account for planning interim analyses.

This article focused on the time for interim analysis for which the ASN is minimized at a fixed power. As for sample size adjustment, other factors are important for selecting the time for interim analyses. Two such factors are the precision in the interim estimates of nuisance parameters and the upper and lower limits of the adjusted sample size. We showed the results when the adjustment did not allow the reduction of sample size from the pre-planned size, because this may cause critical problems from a regulatory viewpoint. Jahn-Eimermacher and Hommel (2007) examined the time for interim analysis when the sample size adjustment included a reduction from the pre-planned sample size. Their results showed a late interim analysis was inferior to a design with a conventional group sequential design having a fixed sample size. The time for sample size adjustment needs to be examined under practical conditions in each trial. The ASN equations provided here can be applied for increasing or decreasing the sample size and limiting any upper/lower sample size; therefore, these can be applied to

the trial's conditions.

3 SAMPLE SIZE RE-ESTIMATION FOR SURVIVAL DATA IN CLINICAL TRIALS WITH AN ADAPTIVE DESIGN

3.1 INTRODUCTION

In clinical trials, adaptive designs have been widely used over the past decade. Among the various adaptive designs, the design used to re-estimate the sample size on the basis of the observed treatment effect have been controversial form various viewpoints such as complicated inferential decisions, the possibility of resulting in clinically meaningless differences, and efficiency (Shi, 2001; Shih, 2006; Taiatis, 2003). Nevertheless, the sample size re-estimation itself has attracted the attention of many investigators, since it is common that uncertainties remain in the critical assumptions about the effect size and extent of data variation during the design of a trial before its start. In the case of continuous and binary data, the methodology for sample size re-estimation has been intensively discussed for two decades. Proschan and Hunsberger (1995) proposed a conditional power based on the treatment effect in an interim analysis and proposed critical values to preserve the type I error rate for the test of a two-sample mean. Cui, Hung and Wang (1999) proposed a test statistic for testing the two-sample means that preserve the type I error rate, using the ordinary critical values of fixed group sequential designs.

Clinical trials to compare the time to events such as death or heart failure tend to be large in size and very long in duration, and they contain fields with a great need for the sample size re-estimation based on interim analyses. Since group sequential designs to terminate a trial early because of futility or success are frequently employed in trials with survival data, the design for re-estimating the sample size should allow the early termination of the trial.

There are various methods for using the survival data to control the inflation of the type I error rate when the sample size is modified on the basis of the observed effect size, see for example Schäfer and Müller (2001); Shen and Cai (2003); Li, Shih and Wang (2005); and Desseaux and Porcher (2007). These procedures are sometimes complicated to be used in actual clinical trials. On the other hand, the procedure of Cui *et al*. (1999) for testing two-sample means seems simple because it uses ordinary critical values in fixed group sequential designs. It would therefore be beneficial to use this procedure for the analysis of survival data. In this article, we employ a log-rank test statistic to which the method of Cui *et al*. (1999) is applied, and investigate the

performance of the test statistic from simulation studies.

In addition to inflation of the type I error rate, the method used to calculate the sample size in an interim analysis should be carefully considered because the data in each stage are mutually dependent in trials with survival data. This article deals with methods to calculate the sample size as well as methods to calculate the target number of events, which is applicable in various practical trials. The hazard ratio estimated from the observed data is sometimes considerably higher than the hypothesized hazard ratio, but is not above the criterion for early termination. If the hypothesized hazard ratio is determined based on some evidence, the interim data may give an incorrect estimate. One solution for such cases is to balance the hypothesized and observed estimations. We propose a method to estimate hazards in order to re-estimate the sample size from this viewpoint. The sample size re-estimation is applied not only to internal pilot or seamless phase 2/3 studies but also pivotal confirmatory studies. The latter is more suitable to the proposed hazard estimation because the confirmatory studies are generally planned based on evidence such as the historical data of clinical studies for the same population as the planning study. For example, the standard error of the hazard ratio in the historical studies and interim results of the ongoing study can be utilized to weight the hypothesized and observed estimation in the proposed method. In addition to the hazard estimation, the shape parameter of survival distribution also has an effect on the sample size re-estimation when the survival distribution is the Weibull distribution. In the planning stage, the exponential distribution tends to be used as the survival distribution due to the lack of information. However, given the interim data, it is easier to estimate the shape parameter. We mention the effect of the shape parameter for the Weibull distribution on the sample size required to achieve the target number of events.

In Section 3.2, we provide an outline of the clinical trial that we deal with in this study. One interim analysis is planned in order to re-estimate the sample size as well as to decide whether or not an early termination for futility or success will be made. We present the test statistic of Cui *et al*. (1999) applied to the log-rank test and other methods to preserve the type I error rate using the log-rank test. In Section 3.3, we describe a method to calculate the minimum number of events required in the second stage. Then, we present a method for calculating the sample size in order to observe the target number of events within the anticipated period. Furthermore, we propose a method to estimate hazards; this method is used to calculate the number of events and sample size in the interim analysis. Section 3.4 presents simulation results which demonstrate the property of the method proposed in Section 3.3. The test statistics

described in Section 3.2 is also compared. In Section 3.5, we present an example using the results of an actual trial in order to show how the proposed method is applied. Finally, in Section 3.6, we briefly discuss some issues that accompany the sample size re-estimation, including the upper/lower bounds of the sample size modified in the interim analysis.

3.2 GROUP-SEQUENTIAL TRIALS WITH SAMPLE SIZE RE-ESTIMATION

3.2.1 Assumptions

We consider a two-arm clinical trial to compare an investigational treatment with a control, in which the endpoint is the time to event occurred first. Let X denote an indicator variable, i.e., $X = 0$ for the control and $X = 1$ for the investigational treatment. The hazard function can be expressed as

$$h(t \mid X) = h_0(t)\exp(-\theta X) \tag{3-1}$$

where $t \geq 0$, $h_0(t)$ is the hazard for the control group, and the hazard ratio is $\exp(-\theta)$. The null hypothesis of no treatment effect is thus expressed as

$$H_0: \theta = 0 \quad \text{vs.} \quad H_1: \theta > 0.$$

Suppose that we plan an interim analysis to decide whether to terminate the trial early for success or futility or to re-estimate the sample size for the second stage and continue the trial. Let z_1 be a certain test statistic to test the treatment effect on the basis of n_1 observations in the interim analysis, and let π_1 be the p-value corresponding to the z_1. Then, for the prescribed threshold values α_0 and α_1, we set the decision rules for the interim analysis as follows:

(A) if $\pi_1 < \alpha_1$, then reject H_0 (early termination for success),

(B) if $\pi_1 \geq \alpha_0$, then accept H_0 (early termination for futility),

(C) if $\alpha_1 \leq \pi_1 < \alpha_0$, continue the trial and enroll n_2 patients in the second stage.

We assume that the interim analysis is performed after d_1 events are observed. Let d_2 denote the preplanned numbers of events to be observed in the second stage of the trial. The target number of events, D (= $d_1 + d_2$), will be estimated at the outset so that the power of the test is over $1 - \beta_1$, under a pre-chosen alternative hypothesis. In the interim analysis, the minimum number of events required after the interim analysis is re-estimated, and we denote it by d_2^*. Thus, the final analysis will be performed when D^* (= $d_1 + d_2^*$) events are observed. Figure 3-1 shows a schema of the trial considered

here.

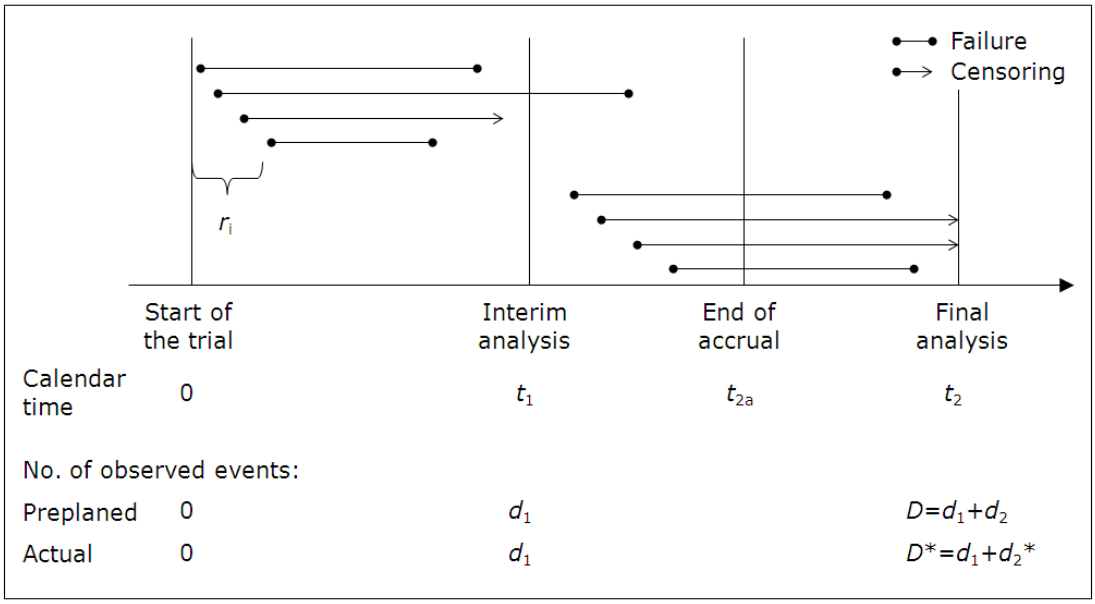


Figure 3-1. Schema of the supposed trial

Although many papers [e.g., Schäfer and Müller (2001), Shen and Cai (2003)] do not require to plan d_2 before the trial starts, most investigators and sponsors need to preplan d_2. The total number of patients is estimated on the basis of the number of events, $D = d_1 + d_2$, which determines the budget and resources for the trial that need to be approved by their organization, and is sometimes required by health authorities and ethics committees. In this article, we assume that the overall sample size is specified at the planning stage prior to the trial.

Let S_j denote the survival function in each group ($j = 0, 1$), and $F_j = 1 - S_j$ be the corresponding distribution function. Suppose that the survival function follows a Weibull distribution with a distribution function $\exp[-(th)^{\gamma}]$, where t is the time from the start of the study, and λ and γ are the parameters of the distribution. The time of the interim analysis is denoted by t_1. The final analysis will be performed at time t_2, and the accrual period in the second stage is $t_1 \leq t \leq t_a$. Let r_i denote the duration from the start of the trial to the time of the enrollment of the ith subject.

3.2.2 Methods to Preserve Type I Error Rate

We deal with the log-rank test, or the Fleming-Harrington Gρ family with $\rho = 0$,

under the assumption of proportional hazard for the simplicity and practicality. Lawrence (2002) showed that changing the value of ρ on the basis of interim results has more power than the log-rank test for non-proportional hazard.

Cui, Hung and Wang (1999) proposed a test statistic, called "CHW statistic" hereafter, as a repeated test statistic for two-sample means or a test statistic following the Brownian motion process, which preserves the type I error rate when the sample size is re-estimated at an interim analysis. The significance level of each test is determined in the same way as a fixed group-sequential design by using an α-spending function (1983). Let $Z(t)$ be the test statistic at the information fraction of t ($0 \le t \le 1$). The standardized version of the test statistic is given by $B(t) = Z(t) \cdot t^{1/2}$, which follows a Brownian motion process (2006). The CHW statistic applied to the log-rank test, and the statistic at the final analysis is

$$\begin{aligned} Z_{CHW} &= Z_1\sqrt{t_1(D)} + \frac{B(t_2(D)) - B(t_1(D))}{\sqrt{t_2(D) - t_1(D)}}\sqrt{1 - t_1(D)} \\ &= Z_1\sqrt{t_1(D)} + Z_2'\sqrt{1 - t_1(D)}, \end{aligned} \tag{3-2}$$

where $Z_2' = \dfrac{B(t_2(D)) - B(t_1(D))}{\sqrt{t_2(D) - t_1(D)}}$, $t_1(D) = d_1/D$, and $t_2(D) = D^*/D$.

Here $t_1(D)$ is the information fraction in the interim analysis which is defined as the ratio between the variance of the log-rank statistics in the interim analysis and the one in the final analysis. The statistic Z_1 indicates the interim test statistic of the log-rank test. Compared with the expression (3-2), the unmodified log-rank test statistic in the final analysis can be expressed as

$$Z = Z_1\sqrt{t_1(D^*)} + Z_2\sqrt{1 - t_1(D^*)},$$

where $Z_2 = \dfrac{B(t_2(D^*)) - B(t_1(D^*))}{\sqrt{t_2(D^*) - t_1(D^*)}}$, $t_1(D^*) = d_1/D^*$, and $t_2(D^*) = D^*/D^*$.

The mathematical properties of the log-rank test show that it follows a Brownian motion process and has independent increments in repeated test: see Schäfer and Müller (2001) and Tsiatis (1981).

Li, Shih and Wang (2005) applies the critical values of Li *et al*. (2002) for continuous data to the survival data. This method adjusts the critical value for the final test to maintain the overall type I error rate of α, given the pre-specified α_0 and α_1. The final critical value, c_I, is calculated by

$$\alpha_0 - \alpha = \int_{z_{1-\alpha_0}}^{z_{1-\alpha_1}} \Phi\left[\frac{c_l\sqrt{d_1+d_2} - z_1\sqrt{d_1}}{\sqrt{d_2}}\right]\varphi(z_1)dz_1 \qquad (3\text{-}3)$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ are the distribution function and the density function of the standard normal distribution, respectively.

Desseaux and Porcher (2007) applies the sample size re-estimation based on Fisher's criterion proposed by Bauer and Köhne (1994) to survival data. Let π_1 and π_2 denote the *p*-values obtained from the interim and final log-rank test, respectively. If $\pi_1 \times \pi_2 < c_{\alpha_2}$ in the final analysis, the null hypothesis is rejected, where c_{α_2} is calculated on the basis of the Fisher's criterion (1948) as

$$c_{\alpha_2} = \exp\left(-\frac{1}{2}\chi^2_{4,1-\alpha_2}\right)$$

where $\chi^2_{4,1-\alpha_2}$ denotes the $(1 - \alpha_2)$-quantile of the chi-square distribution with degrees of freedom of 4.

3.3 NUMBER OF EVENTS AND SAMPLE SIZE REQUIRED IN THE SECOND STAGE

3.3.1 Target Number of Events and Sample Size in the Second Stage

The target number of events for the second stage d_2^* is re-estimated in the interim analysis so that the conditional power becomes greater than $1 - \beta_2$ under the prescribed alternative hypothesis. This can be expressed as $\text{Prob}(Z > z_{1-\alpha_2} | Z_1 = z_1, \theta) \geq 1 - \beta_2$, i.e.,

$$1 - \Phi\left(\frac{z_{1-\alpha_2}\sqrt{d_1+d_2^*} - z_1\sqrt{d_1} - d_2^*\theta/2}{\sqrt{d_2^*}}\right) \geq 1 - \beta_2 \qquad (3\text{-}4)$$

[Proschan and Hunsberger (1995)]. The statistic z_1 is an ordinary log-rank test statistic without any modification as mentioned in Section 3.2.2; it is obtained from the observed data in the interim analysis. Since θ in inequality (3-4) is unknown, the estimated θ^* from the observed data can be used. In the interim analysis, θ^* can be estimated based on the proportional hazard model of equation (3-1) using the method of partial

likelihood (Cox, 1975). This method of estimating θ^* in the interim analysis is the same as that in the final analysis. Although the estimate from the interim data θ^* is commonly used for θ, we propose another approach in Section 3.3.2. The number of events $d_2{}^*$ in inequality (3-4), which we want to re-estimate, cannot be expressed in a closed form and is solved by a numerical calculation. The target number of events based on the unconditional power is given by

$$d_2^* = \frac{4\left(z_{1-\alpha_2} + z_{\beta_2}\right)^2}{\theta_1^{*2}} - d_1. \tag{3-5}$$

Li *et al*. (2005) proposed an equation modified from inequality (3-4), which is consequently the same as equation (3-5). In Section 3.4, we present the effect on the sample size based on the unconditional power instead of the conditional power.

The number of events observed in the second stage consists of the events in subjects enrolled in the first stage and those at risk in the interim analysis, denoted by $S\{R\}$, along with the events in subjects enrolled in the second stage. Under the assumption of a constant accrual rate, we have

$$d_2^* = \sum_{j=0,1}\left[\sum_{i\in S\{R\}} \frac{F_j(t_2 - r_i) - F_j(t_1 - r_i)}{1 - F_j(t_1 - r_i)} + \frac{n_2}{2(t_a - t_1)}\int_{t_1}^{t_a} F_j(t_2 - y)dy\right], \tag{3-6}$$

which is used to calculate n_2.

When the survival function is Weibull, the number of events from $S\{R\}$ on the right-hand side of equation (3-6), which is denoted by d_{2f1}, is given by

$$\begin{aligned} d_{2f1} &= \sum_{j=0,1}\sum_{i\in S\{R\}} \frac{F_j(t_2 - r_i) - F_j(t_1 - r_i)}{1 - F_j(t_1 - r_i)} \\ &= \sum_{j=0,1}\sum_{i\in S\{R\}} \frac{1 - \exp\left(-(t_2 - r_i)^\gamma h_j{}^\gamma\right) - \left\{1 - \exp\left(-(t_1 - r_i)^\gamma h_j{}^\gamma\right)\right\}}{\exp\left(-(t_1 - r_i)^\gamma h_j{}^\gamma\right)} \\ &= \sum_{j=0,1}\sum_{i\in S\{R\}} \left\{1 - \exp\left(-\left((t_2 - r_i)^\gamma - (t_1 - r_i)^\gamma\right)h_j{}^\gamma\right)\right\}. \end{aligned}$$

When $\gamma = 1$ (i.e., an exponential distribution),

$$\begin{aligned} d_{2f1} &= \sum_{j=0,1}\sum_{i\in S\{R\}} \left\{1 - \exp\left(-(t_2 - t_1)h_j\right)\right\} \\ &= \sum_{j=0,1}\sum_{i\in S\{R\}} F_j(t_2 - t_1). \end{aligned} \tag{3-7}$$

Equation (3-7) does not include r_i. Although $\gamma = 1$ is used in general when designing a trial as well as the sample size, the actual hazard may increase ($\gamma > 1$) or decrease ($\gamma < 1$)

over time in the clinical field. If $\gamma = 1$ is supposed against the true survival distribution of the Weibull distribution with $\gamma = 2$, d_{2f1} will be underestimated by the following difference:

$$\begin{aligned}
&\sum_{j=0,1}\sum_{i\in S\{R\}}\left\{1-\exp\left(-(t_2-t_1)h_j\right)\right\}-\sum_{j=0,1}\sum_{i\in S\{R\}}\left\{1-\exp\left(-\left((t_2-r_i)^2-(t_1-r_i)^2\right)h_j{}^2\right)\right\}\\
&=-\sum_{j=0,1}\sum_{i\in S\{R\}}\left\{S_{\gamma=1}(t_2-t_1)F_{\gamma=1}\left((t_2-r_i)+(t_1-r_i)\right)\right\}\\
&<0.
\end{aligned}$$

Consequently, superfluous subjects are enrolled in the second stage of the trial. To mitigate this risk, the interim estimate of γ can be used for calculating $d_2{}^*$, although it should be considered carefully if the interim estimate of γ <1 (or >1) against the anticipation of γ >1 (or <1).

3.3.2 Method to Estimate Hazards

When the conditional power is calculated at the interim analysis, the hazard ratio estimated from the observed data is commonly used. However, if the estimated hazard ratio is considerably higher than the hypothesized hazard ratio but is not above the criterion for early termination, the observed data in the first stage may not represent the population well. In that case, the required number of events will probably exceed the pre-specified upper bound. If the hypothesized hazard ratio is determined based on some evidence, one remedial solution is to balance both sources of information. We propose a method to estimate the hazard ratio using such a view.

Let $S_{aj}(t)$ denote the survival function anticipated before the trial starts and $S_{nj}(t)$ denote the survival function estimated from the observed data in each treatment group (j = 0, 1). Suppose that the survival function follows a Weibull distribution $\exp[-(th)^{\gamma}]$, and also suppose that $S_{aj}(t)$ and $S_{nj}(t)$ have the same parameter γ_j. The difference between them in a log-log scale, η_j, following the notation of Whitehead *et al*. (2001), is given by

$$\begin{aligned}
\eta_j &= -\log\left(-\log S_{nj}(T)\right)+\log\left(-\log S_{aj}(T)\right)\\
&= -\log\left(-\log\exp\left[-(Th_{nj})^{\gamma_j}\right]\right)+\log\left(-\log\exp\left[-(Th_{aj})^{\gamma_j}\right]\right)\\
&= -\gamma_j\log T-\gamma_j\log h_{nj}+\gamma_j\log T+\gamma_j\log h_{aj}\\
&= \gamma_j\log h_{aj}/h_{nj}\,.
\end{aligned}$$

We propose an interim estimation of the survival function $S_j{}^*(t) = \exp[-(th_j{}^*)^{\gamma j}]$ as

$$-\log\left(-\log S_j^{*}(t)\right) = -\log\left(-\log S_{aj}(t)\right) + c\,\eta_j , \qquad (3\text{-}8)$$

where c is a constant within (0, 1). The survival function $S_j^*(t)$ is consistent with $S_{nj}(t)$ if $c = 1$; consistent with $S_{aj}(t)$ if $c = 0$; and intermediate between them if $c = 1/2$. The interim estimate of the hazard is given by $h_j^* = h_{aj}^{1-c}\, h_{nj}^{c}$ from equation (3-8). The hazard h_{nj} from the observed data can be estimated using the method of maximum likelihood. In the example presented in Section 3.5, we estimated the hazard using the statistical package of the SAS LIFEREG procedure. Once h_j^* is estimated, $\theta^* = -\log(h_1^*/h_0^*)$ can be used for θ in inequality (3-4) on which the target number of events for the second stage was re-estimated.

The different parameter γ can be assumed for $S_{aj}(t)$ and $S_{nj}(t)$ as η_j is replaced with $\eta_j(t) = (\gamma_{aj} - \gamma_{nj}) \log t + \gamma_{aj} \log h_{aj} - \gamma_{nj} \log h_{nj}$ in equation (3-8), where γ_{aj} and γ_{nj} are γ_j for $S_{aj}(t)$ and $S_{nj}(t)$ respectively.

3.4 SIMULATION

A simulation study was performed to demonstrate the property of the method proposed in Section 3.3. Before that, comparisons for the test statistics described in Section 3.2.2 and the approach to d_2^* based on conditional and unconditional power are presented.

In the simulation study, random events were generated under the assumptions that the hypothesized hazard ratio was 0.8, the one-year event rate for the control group was 0.2, and the survival function was a Weibull distribution with $\gamma = 1$. The accrual period and the minimum period of follow-up were both two years, and the additional accrual period after the interim analysis was one year. The accrual was performed uniformly in this period. The power to be achieved was 0.9, which did not take into account of the inflation in the type II error rate caused by an early termination for futility because of its negligible effect. The overall significance level of α was 0.025 (one-sided). The constants α_0 and α_1 were determined by the O'Brian-Fleming type of α-spending function, and d_2^* was calculated using the equation presented in Section 3.3.1 as equation (3-6) to compare the overall power among the methods. Here, we set the upper bound for D^* as four times of D, where D was 849. It was not allowed to be decreased from the preplanned sample size in the modification of the sample size.

First, we compare the overall power of the CHW statistic with the methods of Li, Shih and Wang (2005) and one of Desseaux and Porcher (2007) in the simulation study.

The overall power is the probability of accepting the alternative hypothesis in the interim analysis or in the final analysis. All the three methods showed some elasticity in the overall power, while the overall power in the fixed design without sample size re-estimations fell off as the expected hazard ratio differed from the true one (Figure 3-2).

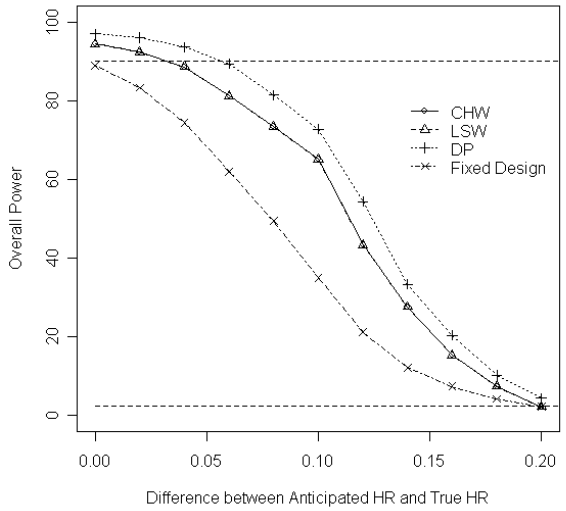


Figure 3-2. Overall power against the difference between anticipated hazard ratio (0.8) and true hazard ratio (0.8 to 1)

CHW: the method applied from Cui, Hung, and Wang (1999)

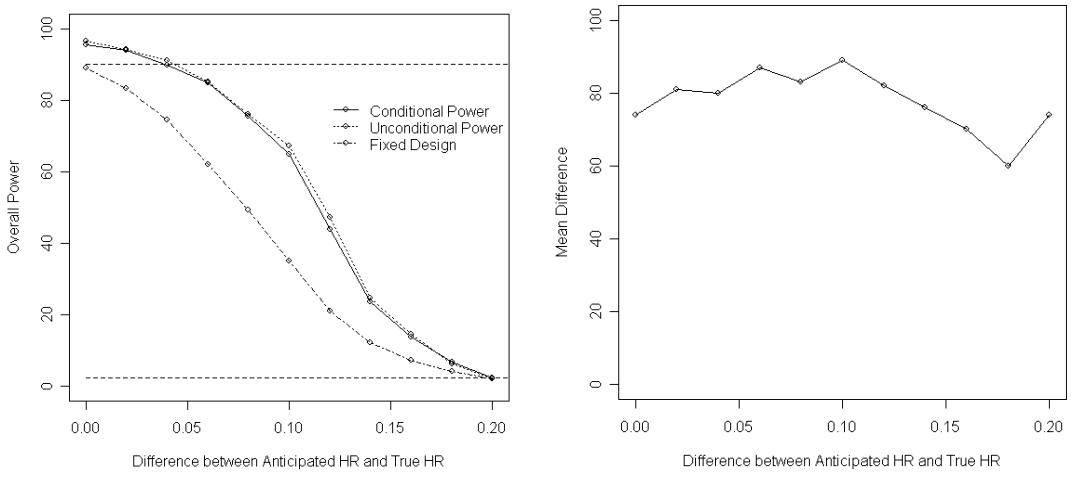
LSW: the method of Li, Shih and Wang (2005)

DP: the method of Desseaux and Porcher (2007)

The overall power under the null hypothesis (i.e. true hazard ratio = 0) was 2.5% for the CHW method and 2.6% for the method of Li, *et al* (2005). In the method of Desseaux and Porcher, the overall power was slightly higher than that in the CHW method. As the results of our simulation, the observed overall power of 3.5% under the null hypothesis was over the target type I error rate of 2.5%, which was consistent with the simulation results of Desseaux and Porcher (2007) since they did not intend for the type I error rate to always be below the nominal level. In conclusion, the CHW method is comparable to the other methods in terms of its overall power and ability to control the type I error rate.

Second, we compare the approach to d_2^* based on the conditional power of the equation (3-4) with the unconditional power of equation (3-5). Figure 3-3(a) shows the

overall power of each approach. The type I error rate was controlled by the CHW method. While the mean number of events required in the second stage (d_2^*) based on the conditional power was less than that based on the unconditional power (see Figure 3-3(b)), the overall power was almost the same in the two methods. As a natural result, the difference in the sample size for the second stage between the conditional power and unconditional power was further increased. Since the enrollment number of subjects is never trivial, it can be concluded that the conditional power is more effective in the calculation of the number of events.



(a) (b)

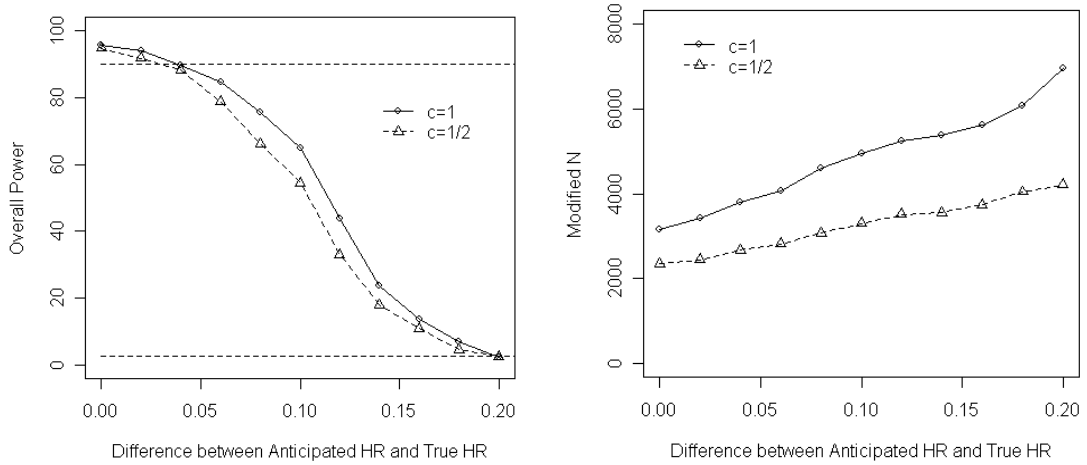
Figure 3-3. Comparison of simulation results between conditional power and unconditional power at the anticipated hazard ratio = 0.8

(a) Overall powers for the sample size modified at the interim analysis based on the conditional power and based on the unconditional power, and overall power for the fixed design (without sample size modification)

(b) Differences between the mean number of events required in the second stage (d_2^*) based on the conditional power and that based on the unconditional power (d_2^* based on unconditional power - d_2^* based on conditional power)

Finally, we applied the hazard estimation based on equation (3-8) to re-estimate the sample size in the simulation study. The sample size was re-estimated based on the conditional power. The overall power and mean sample size were compared using equation (3-8) with c = 1 (i.e., the hazard estimated from the observed data in the first stage), 3/4, 1/2 (i.e., the intermediate between the hypothesized and the observed hazards), and 1/4 (Figure 3-4). As c was larger, the overall power and mean sample size

were larger. However, the increase in mean sample size was remarkably large compared with the increase in overall power. For example, when the difference between the hypothesized and true hazard ratios was 0.1, the overall power with c = 1 increased by 12% compared to that with c = 1/2, but the mean sample size with c = 1 increased by 1658. When there was no difference between the hypothesized and true hazard ratios, the overall power with c = 1 increased by 1% compared to that with c = 1/2, but the mean sample size with c = 1 increased by 870. In addition, the percentage of cases where the re-estimated sample size was greater than the upper bound was 10% to 33% when c = 1, while the percentage was zero when c = 1/2. Investigators and sponsors could choose c to be less than 1 to avoid the considerable increase in the sample size, in exchange for some loss in the overall power. We provide an example of setting c to less than 1 in the next section.



(a) (b)

Figure 3-4. Comparison of simulation results between sample size modification based on the hazard estimated by the equation (3-7) with c = 1 and with c = 1/2 at the anticipated hazard ratio = 0.8

Overall powers for the sample size modified at the interim analysis based on c = 1 and based on c = 1/2

Mean sample sizes modified at the interim analysis based on c = 1 and based on c = 1/2

3.5 EXAMPLE

The Eplerenone Post-Acute Myocardial Infarction Heart Failure Efficacy and Survival Study (EPHESUS) was a double-blind, placebo-controlled clinical study evaluating the effect of eplerenone, a selective aldosterone blocker, on morbidity and mortality among patients with acute myocardial infarction complicated by left ventricular dysfunction and heart failure (Pitt, *et al*., 2003). The co-primary endpoints were the time to death from any cause and the time to death from cardiovascular causes or the first hospitalization for a cardiovascular event. The trial was designed to enroll 6200 patients and to continue until 1012 deaths occurred. The target number of deaths was determined to have an 88% power to detect an 18.5 % reduction from the placebo in the hazard of death using a log-rank test at a significance level of 0.04. As the results, 478 of the 3319 patients in the eplerenone group and 554 of the 3313 patients in the placebo group died. The hazard ratio for death was 0.85 (95% confidence interval, 0.75 to 0.96). The difference in the hazard ratio between the anticipation and result was 0.035. While the power to detect the reduction in the hazard ratio was overestimated, the reduction was statistically significant because of the original high power of 88%.

Here, we suppose that the hazard ratio of 0.8 was anticipated against the true hazard ratio of 0.85. The target number of deaths D is 897 to achieve a power of 90% under the hypothesized hazard ratio, where an interim analysis for early termination for futility or success is planned after 50% of the preplanned number of deaths have occurred. With 897 deaths, the probability of failing to detect a significant reduction in the hazard ratio is 36% (i.e., 64% power) under the true hazard ratio if no sample size modification is performed. We suppose that sample size re-estimation is planned in the interim analysis in addition to the early termination.

We conducted a simulation study based on the above conditions. In order to determine c in equation (3-8), the standard error of the hazard ratio in a historical study and the interim results of the ongoing study were used. The historical study used was the TRACE study [20], which evaluated the effect of an ACE inhibitor against a placebo in patients with left ventricular dysfunction after myocardial infarction, resulted in a hazard ratio of 0.78 (95% confidence interval, 0.67 to 0.91) in the event of death. The hazard and its standard error based on the observed data in the first stage were estimated by the SAS LIFEREG procedure. When c was based on the inverse standard error of the hazard ratio in the interim analysis and TRACE study, it was approximately 1/2 in the simulation study. Table 3-1 shows the estimates from the observed data and

pre-specified values in the simulation study, which were used to calculate the weighted hazard ratio with c in equation (3-8). This weighted hazard was used for θ in inequality (3-4), on which the target number of events for the second stage d_2^* was re-estimated. The log-rank test statistic z_1 in inequality (3-4) was calculated by the SAS LIFETEST procedure from the observed data in the first stage. The minimum d_2^* that satisfied inequality (3-4) with 0.9 of 1 - β_2 was obtained through a numerical calculation. The number of subjects enrolled in the second stage was then calculated by equation (3-6) with an upper bound equal to four times D.

Table 3-1. Estimates in interim analysis and pre-specified values relative to weighted hazard ratio with *c*

Estimated or pre-specified value	Mean† (min, max)
Standard error of hazard ratio from observed data	1.099 (1.098, 1.102)
Standard error of hazard ratio in the historical study	1.08*
c	0.495 (0.495, 0.496)
Hazard ratio from observed data	0.853 (0.592, 1.200)
Hypothesized hazard ratio	0.80*
Weighted hazard ratio with c	0.825 (0.689, 0.978)

†: Mean of estimates of 10,000 simulated studies, except for pre-specified values

*: Pre-specified value

In the results of the simulation study, the overall power was 80%, and the mean D^* was 1279. When the sample size re-estimation was performed with c = 1, the overall power under the true hazard ratio was 86%, and the mean D^* was 1800. This means that the value of D^* when c = 1/2 was lower than when c = 1; this was despite the overall power being greater than 80%, and it is even lower than the D of 1692 when the hypothesized hazard ratio was equal to the true one of 0.85.

3.6 DISCUSSION

In the proposed method to estimate interim hazards, constant c is set in proportion to the degree of confidence in the hypothesized hazard and the observed data. We would recommend specifying the decision rule of the c at the planning stage, while it may be difficult to pre-specify the value of c itself. The information fraction of the interim analysis can be one possible factor to determine the value of c. The proposed method

with c = 1/2 could control excessive increase of sample size modified at the interim analysis, while little reduction in the overall power was observed. However, if investigators are highly confident in the hypothesized hazard ratio, there is no need for the sample size re-estimation, and the fixed design should be employed.

The CHW method applied to the log-rank test was comparable to other methods in terms of the overall power and control of the type I error rate. Since the CHW method uses α for the fixed design, it can easily be applied in actual clinical trials. However, this may cause critical problems from a regulatory view point if the sample size re-estimation allows the sample size to be reduced from the preplanned sample size. If the interim analysis does not result in an early termination for success but results in a modified number of events equal to zero or a very few, there may be serious doubt about leaving the final significance level as the preplanned α_2 in the CHW method. Cui *et al*. (1999) does not suppose a decrease in the sample size at the interim analysis. If the CHW method is used, the sample size may not be reduced from the preplanned one at the interim analysis.

In the simulation discussed in this article, we set an upper bound on the additional number of events, d_2^*, in the second stage of the trial, and did not allowed a decrease in the sample size when modifying it. The upper bounds for d_2^* and n_2 are mandatory, because n_2 can be tens times as large as the preplanned size, even if the study does not terminate for futility. The upper bound should be determined so that it is at least less than the size needed to detect a clinical meaningful difference. The budget of the study and feasibility may further decrease this upper bound.

The lower bounds for d_2^* and n_2 are more complicated. When the sample size can be reduced at the interim analysis, the time of the interim analysis or the accrual duration should be planned so that the interim analysis is performed before the end of enrollment. In the results of a simulation, the percentage of such cases was low when the interim analysis was conducted at the information fraction of 1/2. As the information fraction became smaller than 1/2, the percentage of the completion of the enrollment was getting lower. However, the percentage of the early termination for success was also low, which made the mean sample size larger. Furthermore, as mentioned above, if the interim analysis results in a modified number of events equal to zero or a very few, the final significance level of α_2 reaches an *impasse* in the CHW method. Therefore, we think that it is practical for any decrease in sample size at the interim analysis to be caused only by an early termination for futility or success, while any modification involves an increase.

4 CLINICALLY IMPORTANT EFFECTS IN NEW DRUG DEVELOPMENT

4.1 INTRODUCTION

In new drug development, demonstrating the clinical importance of the drug effect should be a key element of efficacy evaluations in addition to simply showing a statistical significance. The clinical importance or clinical meaningfulness of the new investigational drug is defined and assessed generally by comparing drug effects with the Minimal Clinically Important Change from baseline (MCIC) or Minimal Clinically Important Difference between groups (MCID) of a primary endpoint. While the clinical importance of drug effect is often evaluated when interpreting results of clinical trials, it is also important to consider the clinical importance at the planning phase. In recent years, adaptive designs have been widely used in the field of clinical trials, and many trials have employed the Sample Size Re-estimation (SSR) during the trial. When the SSR is implemented, it is recommended to set the upper limit of the sample size based on the clinical importance (Shih, 2001; Hung *et al*., 2006). The concept of clinical importance is also applied to the non-inferiority margin in non-inferiority trials (D'Agostino and Massaro, 2003).

What is the widely accepted definition of the clinically important effect in the first place? There are various definitions, such as: changes in restoring normal levels of functions by the end of drug therapy, changes in significantly reducing patients' risk for various health problems, changes in a level of what is recognized and accepted as "improvement" by doctors and patients, and so on (Jacobson and Truax, 1991). While the definitions of clinical importance have been discussed by clinicians for many years, sometimes those discussions might confuse MCIC and MCID. The confusion of MCIC and MCID may cause incorrect use, for example, if the treatment difference is compared with the MCIC.

There are some disease areas such as hypertension and chronic pain where the MCIC or MCID is well investigated. However, the MCIC and MCID are considered differently among diseases and endpoints. Also, the MCIC and MCID sometimes change with the improvement of standard treatments. When there is no MCIC or MCID established in a disease area, one approach is to conduct a survey of well-experienced clinicians on how much effect they would consider the MCIC or MCID. In other cases, the MCIC or MCID is derived from the results of historical trials for the standard treatment. Some common approaches can be taken for defining the MCIC and MCID in

various disease areas.

Even if the MCIC or MCID is determined, there are remaining points to be considered; how to compare the trial results with the MCIC or MCID, and how to decide whether the effect of the new drug is clinically important from the comparison. For example, if the mean is slightly lower than the MCID, should we judge that the drug doesn't have a clinically important effect? The probability of the mean being greater than the MCID is only 50% under the hypothesized mean which equals the MCID.

In this article, we will provide the concept of MCIC and MCID and make clear the way to use them in new drug development through resolving the issues or uncertainties relative to MCIC and MCID aforementioned. In the next section, we present the distinction between MCIC and MCID. Section 4.3 presents each role of MCIC and MCID in PoC trials and superiority trials. In Section 4.4, three representative approaches to estimate the MCIC and MCID are reviewed. Furthermore, we would provide general approaches to interpret the trial results compared with the MCIC or MCID in Section 4.5.

4.2 MCIC AND MCID

While how much each drug improves the individual patient's health problem is a crucial aspect in clinical practice, it is important for new drug development to assess whether the investigational drug has a clinically important effect for the population of the target disease. In order to demonstrate clinical importance, there are two representative criteria mentioned in the introduction: MCIC of the minimal clinically important change from baseline of pre-dosing to a certain time point in a primary endpoint; and MCID of the minimal clinically important difference between treatment groups.

For describing MCIC, we first focus on an observation of an individual subject in the change from baseline. The individual change (continuous data) is evaluated in terms of whether it achieved the MCIC, i.e. being less or more than the MCIC as binary data (Figure 4-1). Furthermore, the percentage of subjects who achieved the MCIC in the total population is calculated. Sometimes this numerical value of the percentage can provide a clinical meaning, and, in other cases, the percentage has a clinical meaning only when compared with the control.

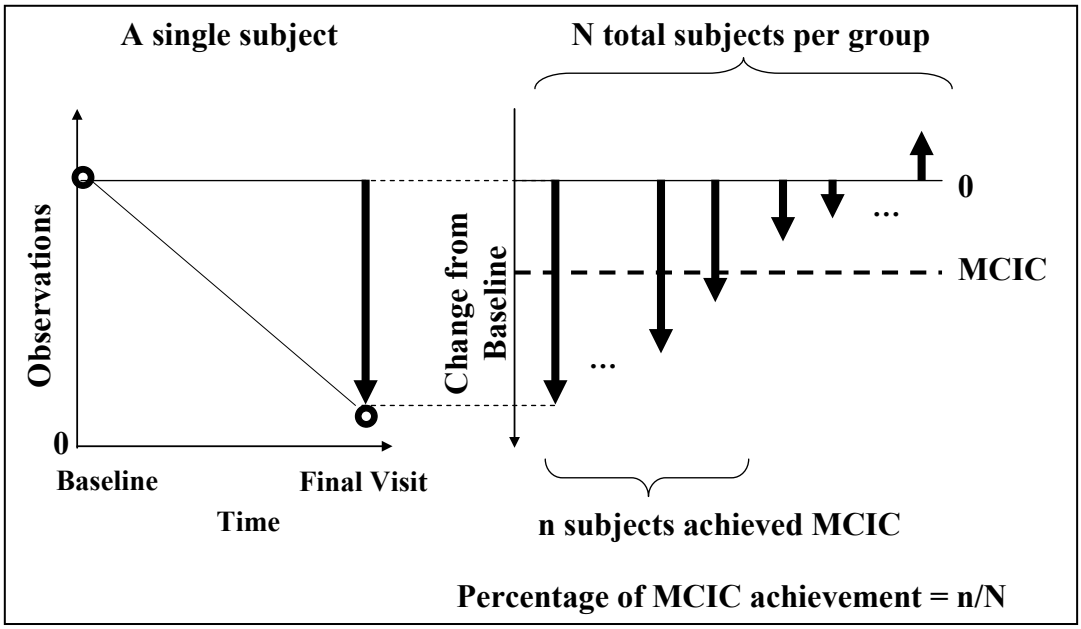


Figure 4-1. Individual change from baseline and MCIC

The MCID is generally used for the comparison between treatment groups of the investigational drug and the control; it is not used for the assessment of individual subjects. For example, a treatment difference in the mean change from baseline is estimated for each treatment group, and whether or not a treatment difference is beyond the MCID is evaluated (Figure 4-2).

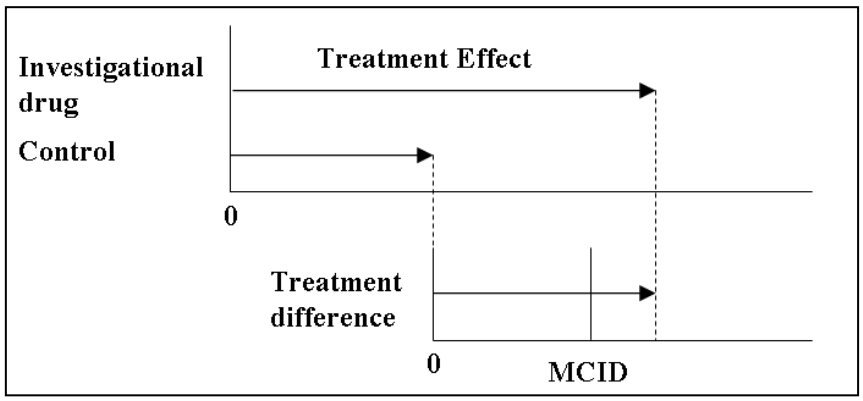


Figure 4-2. Treatment effect and MCID

4.3 ROLES OF MCIC AND MCID IN CLINICAL TRIALS

Situations where the MCIC and MCID play a key role are: 1) screening of investigational drugs, 2) estimating sample size, 3) re-estimating sample size during the trial, and 4) interpreting results of trials. We describe the details of each role in two types of clinical trials.

- **'Proof of Concept' Trial**

In Proof of Concept (PoC) studies, investigational drugs are screened for late phase developments based on preliminary evidence of safety and efficacy. The decision criteria for PoC must generally be customized for every mechanism of action (Cartwright, *et al*., 2010). As for efficacy, what we should assess is whether the results of the PoC trial indicate a sufficient probability to show the clinically important effect in future confirmatory studies. It is important to compare the confidence interval of the mean change from baseline in the investigational drug and MCIC. If the trial has the control of placebo or the standard drug, the confidence interval of the treatment difference from the control can be compared with MCID. The confidence level of the confidence interval depends on the acceptable decision error rates for false-positive and false-negative results.

- **Superiority Trial**

In superiority trials, sponsors and investigators want to conclude that the treatment difference is larger than MCID besides the statistical significant difference. Some textbooks state that the sample size should have a desired power to detect the MCID on the primary endpoint in superiority trials (e.g., Chow and Liu, 2004). On the other hand, if the sample size is determined based on the MCID, in spite of anticipating the far larger treatment difference than the MCID, it is inefficient in terms of demonstrating the statistical significance in the treatment difference (Golub, 2006). Therefore, the anticipated effect size is sometimes used in the sample size determination while there is a risk that investigators and sponsors tend to over-expect the drug effect.

Recently, many trials have employed the SSR because of the uncertainty of the anticipated effect size. In the SSR, the upper limit of the re-estimated sample size is necessary. The factors determining the upper limit are not only the trial cost and feasibility but also the MCID because the huge sample size needed to detect clinically meaningless difference is unethical and troublesome (Shih, 2001; Hung, *et al*., 2006). A

conventional group sequential design is also effective when the sample size is pre-planned to have adequate power to detect the MCID, and the trial can be terminated early with success (Jennison and Turnbull, 2006). In either way, the MCID is a key point of the trial design.

4.4 APPROACHES TO MCIC AND MCID

There are three representative approaches to estimate MCIC and MCID: 1) Distribution-based approach, 2) Anchor-based approach, and 3) Opinion-based approach.

4.4.1 DISTRIBUTION-BASED APPROACH

The distribution-based approach is based on the comparison of two different distributions of an endpoint. For the two distributions, we can assume: i) healthy/normal or patient/abnormal distribution (i.e. functional or dysfunctional population), ii) before or after treatment distribution, and iii) control or investigational drug group distribution. For i) the functional and dysfunctional distributions, Jacobson, et al. (1991) defined clinically important change as something to do with the return of normal functioning. While it well represents the view of clinicians, this change doesn't mean the size of the change from baseline and doesn't connect with the MCIC.

In the case of ii) before and after treatment distributions and the case of iii) distributions of control and investigational drug groups, some methods relative to the effect size or standard error of measurement have been proposed. Cohen's *d* (1988) is one common method among them. The *d* is effect size, i.e. a difference of two means divided by the standard deviation. Cohen provided a common conventional frame to interpret *d* with the *proviso* that it's relative to the area or research method: 0.20 for "small" effects, 0.50 for "moderate" effects, and 0.80 for "large" effects. These correspond to areas of two distributions which do not overlap are: 14.7% when $d = 0.2$, 33% when $d = 0.5$, and 47.4% when $d = 0.8$ (shaded area in Figure 4-3). This common frame could be adopted into MCIC when we are interested in the distributions of before and after treatment, and into MCID when we are interested the distributions of the control and investigational drug groups.

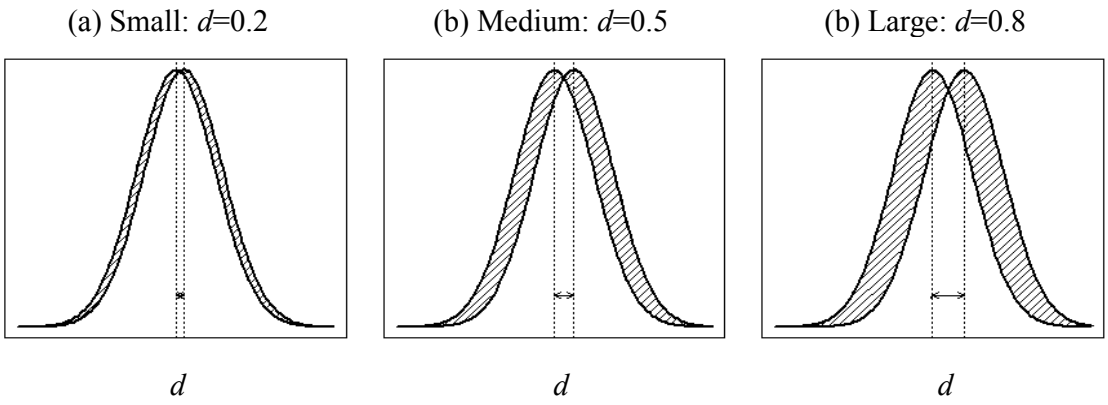


Figure 4-3. Cohen's *d* and areas of two distributions which do not overlap

4.4.2 ANCHOR-BASED APPROACH

The anchor-based approach compares the measure of interest to other standard outcome measures, preferably those that have importance with more universal agreement and those with established MCIC (Make, 2007; Sierevelt, *et al*., 2007). We can explain this approach using an example. Farrar, et al. (2001) assessed the MCIC for pain intensity measured by Pain Intensity Numerical Rating Scale (PI-NRS) based on placebo data from ten clinical trials. Patient Global Impression of Change (PGIC), which was the standard outcome, was closely correlated with the PI-NRS. The PGIC's categories of "much improved" and "very much improved" were considered as MCIC and used as determination of MCIC for the PI-NRS. The average change from baseline in the PI-NRS (a reduction of approximately 30%) corresponding to those two categories in the PGIC could be the MCIC for the PI-NRS.

4.4.3 OPINION-BASED APPROACH

The opinion-based approach gathers the opinions of experts, patients, and health-care practitioners (Sierevelt, *et al*., 2007). It is common to survey opinions from physicians who are Key Opinion Leaders in the disease area. If the test or scale used in clinical trials is used in clinical practice, physicians evaluate many patients using the test and have a feeling of how large the clinical importance is. On the other hand, even if the physician is well experienced and the leader of the area, it would be better to survey many physicians because opinions are subjective. In addition, surveys of

physicians selected from one demographic area, one hospital, or one small study group may produce biased results. When MCID based on the opinion-based approach is used for the sample size determination of confirmatory studies, it's important to confirm the regulatory acceptance.

In some disease areas, surveys using the opinion-based approach were conducted. For example, Burback, et al. (1999) analyzed survey results from 161 physicians in Canada. The mean of the MCIC in Mini-Mental State Examination for dementia patients was 3.72 based on their opinions.

4.5 INTERPRETATION OF TRIAL RESULTS FOR CLINICAL IMPORTANCE

Once the MCIC or MCID is determined, the next step is how to interpret whether trial results indicate that the investigational drug has the clinically important effect based on the MCIC or MCID. Here, we presume that the statistical significance in the primary analysis of the trial has been shown then the clinical importance of the drug effect is discussed.

As MCIC is determined, the percentage of subjects who achieved the MCIC can be an endpoint. These subjects are sometimes called "responders". For the percentage considered as the clinical importance, a high value may be required in some disease areas, but in progressive disease areas without any standard treatments, a low percentage over the measurement error can be acceptable. If the value of the percentage alone does not give such interpretation, the difference between the percentages of the investigational and control drugs is compared with MCID.

After the MCID is determined, the clinical importance can be demonstrated by showing the mean difference between the investigational drug and the control being greater than the MCID. However, it is still controversial as to whether the point estimate of the treatment difference or the lower or upper limit of the confidence interval (CI) should be used. Man-Son-Hing, et al. (2002) suggested the following four forms for the extent of the clinically important effect which the investigational drug has (see also Figure 4-4):

A. definite – when the MCID is smaller than the lower limit of the 95% CI;
B. probable – when the MCID is greater than the lower limit of the 95% CI, but smaller than the point estimate of the mean treatment difference;
C. possible – when the MCID is less than the upper limit of the 95% CI, but greater than the point estimate of the mean treatment difference; and
D. definitely not – when the MCID is greater than the upper limit of the 95% CI.

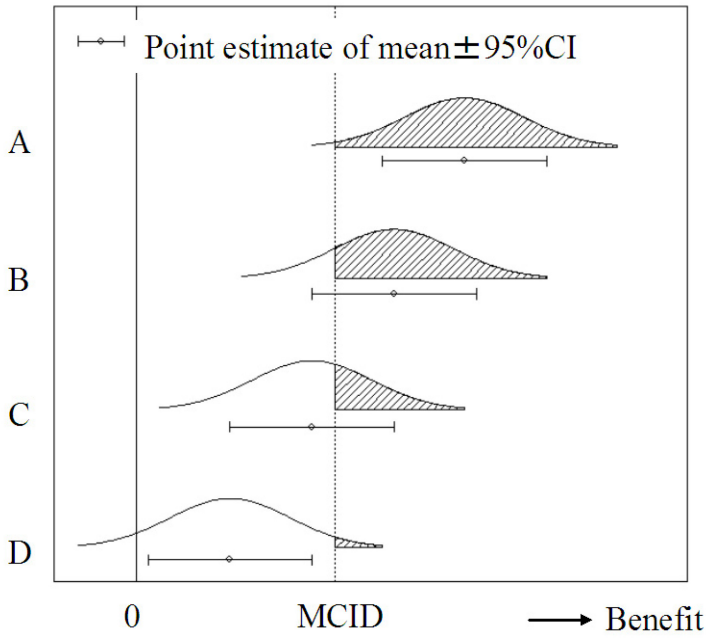


Figure 4-4. Relationship of MCID to point estimate of mean treatment difference and 95% confidence interval (CI)

Given the statistical significance in the treatment difference, we consider the clinical importance is demonstrated by showing the point estimate being greater than the MCID. Even if the point estimate is a little smaller than MCID, it is important to evaluate the probability of the mean treatment difference being greater than MCID based on the distribution of the mean (i.e. diagonally hatched area in Figure 4-4C). However, these evaluations using MCID are reliable only when the sample size has a high power to detect the treatment difference that is greater than MCID. Otherwise, one cannot deny that the results of the mean treatment difference being greater than MCID occur by chance. Therefore, the discussion of the MCID at the planning stage is

important for superiority trials.

4.6 DISCUSSION AND CONCLUSION

In this article, we stated the distinction between MCIC and MCID, and their roles in new drug development. Furthermore, we presented how the clinically important effect of the drug should be demonstrated using MCIC or MCID in clinical development. We would recommend discussing the clinically important effect at the planning phase of trials. The MCIC and MCID may sometimes change because the results of PoC trials may reveal new profiles of the drug or the standard treatment may improve during the drug development. Hence, investigators and sponsors should keep reviewing the MCIC and MCID even after those are established.

Although this article focused on the clinical importance of efficacy, it is also necessary to evaluate safety and the risk-benefit. In disease areas such as oncology where the safety risk is high, the evaluation of clinical importance of the efficacy alone can be insufficient because the acceptable risk and risk-benefit depends on individual patients. In addition when the efficacy endpoint is survival time or disease-free survival time, it is difficult to define how long prolongation is clinically important due to various patient's values of the Quality of Life. In such cases, pharmacoeconomics is one of the best methods to evaluate the value of the drug, comparing the cost of medication for an additional 1-year survival or quality-adjusted survival, which patients or their family can accept, with the required cost of medication including the new drug.

We hope that this article will help more people use the idea of "clinically important effect" for planning of clinical trials or entire development programs for new drugs. We consider this will facilitate more informative clinical trials that provide valuable means to patients and doctors.

5 GROUP COMPARISONS INVOLVING ZERO-INFLATED COUNT DATA IN CLINICAL TRIALS

5.1 Introduction

In clinical trials, outcomes of count data sometimes have excess zeros. These outcomes may be found in the number of symptoms (e.g., urinary incontinence episodes, gastrointestinal ulcers, hot flushes arising from menopausal disorder), the number of events (e.g., hospitalizations, heart attacks), and questionnaire scores. When zero values are observed after treatment in clinical trials to evaluate treatment effect, it represents two types of outcome: that in which patients recover, and that in which patients do not recover but have a small value of the outcome that is zero by chance. The zero-inflated Poisson (ZIP) distribution or zero-inflated negative binomial distribution can be applied to count data with excess zeros. This article focuses on the ZIP distribution. The ZIP distribution has two parameters; λ is the Poisson parameter and ω expresses the extent of zero-inflation compared with zero counts that occur from the Poisson distribution.

When the treatment difference between a test drug and a control is tested, either zero-inflation or the difference in the non-zero part is sometimes ignored. For example, a nonparametric test may be used after applying rank transformation to the data. This poses the problem that there can be many ties due to zero-inflation. In this case, the normal approximation is not accurate, whereas many nonparametric tests use the normal approximation for rank data. In addition, the power to detect the treatment difference could be decreased if there are many ties in the two treatment groups. Another example is that the treatment groups are compared only in the proportion of subjects with zero as a dichotomous response. This possibly wastes the treatment difference in the non-zero part by ignoring that.

In contrast, there are two methods to compare treatment groups considering zero-inflated counts. One method is the zero-inflated count model (Lambert, 1992), and the other is the two-part model (Heilbron, 1992). The zero-inflated count model is a mixture model of a logit model for ω and a Poisson regression model for λ. The treatment effect is assessed by including a covariate for the treatment group in the logit and Poisson regression models. In clinical trials, however, it may be undesirable to assess the treatment effect by performing two hypothesis tests for each model because of the difficulty in interpreting the two possibly inconsistent test results and controlling the type I error inflation.

The two parts of the two-part model are defined as the zero part, consisting of the response dichotomized as zero vs. non-zero, and the non-zero part, consisting of the non-zero counts. The response variable follows the binominal distribution in the zero part and the zero-truncated count distribution in the non-zero part. By applying the two-part model, Lachenbruch (2001a) proposed a test statistic called the two-part statistic that combines the test statistics of the zero and non-zero parts. The test for the zero part is generally the chi-square test. Possible tests for the non-zero part are the Wilcoxon test, *t*-test, etc. This two-part test statistic provides a comprehensive result for both the zero and non-zero parts. Delucchi et al. (2004) demonstrated the two-part statistic with zero-inflated counts in the Addiction Severity Index in heroin addicts. As an alternative to the two-part statistic, Hallstorm (2010) proposed a modified Wilcoxon test for zero-inflated data that discards an equal number of zeros in each group.

This article provides methods for finding the sample size and power for the two-part statistic. Lachenbruch (2001b) developed a method for calculating the sample size when the test for the non-zero part was a *t*-test. Given the considerable ties in count data, however, Lanchenbruch's method can underestimate the sample size when the Wilcoxon test is used for the non-zero part. We propose methods that, by adjusting for ties, calculate the sample size and power for the two-part statistic when the Wilcoxon test is used for the non-zero part. Furthermore, we examine, under three cases, the power of the two-part statistic compared with the conventional methods and the ZIP model. There is treatment difference in the zero-inflation ω under the first of these cases, treatment difference in the Poisson parameter λ under the second, and treatment difference in both parameters under the third. Section 5.2 presents the methods for calculating the sample size and the power using the two-part statistic. The relationship between the non-zero part and the zero-truncated Poisson distribution is also described. In Section 5.3, the power of the two-part statistic introduced in Section 5.2 is compared with the conventional methods. Section 5.3 also describes the ZIP model for assessing the treatment effect, and includes a comparison of the two-part statistic and the ZIP model using a simulation study. In Section 5.4, an application of the proposed method for calculating the sample size is illustrated in an example.

5.2 Two-part statistics and sample size

Consider a clinical trial to compare a test drug with a control. In the trial, N subjects are allocated to the treatment groups in the ratio r: $1 - r$ ($0 < r < 1$). A response

variable for the primary endpoint Y_i follows the ZIP distribution that is given by

$$P(Y_i) = \begin{cases} \omega_i + (1-\omega_i)p_C(0) & (Y_i = 0) \\ (1-\omega_i)p_C(Y_i) & (Y_i > 0) \end{cases} \tag{5-1}$$

$$p_C(Y_i) = \lambda_i^{Y_i} e^{-\lambda_i} / Y_i!.$$

Here, ω_i and λ_i are unknown parameters where $0 < \omega_i < 1$ and $\lambda_i > 0$. The case of the zero-deflated distribution $\omega_i < 0$ is not covered in this article. The null hypothesis is

$$H_0: p_1 = p_2 \text{ and } \mu_1 = \mu_2,$$

where p_i represents the proportion of zero counts

$$\begin{aligned} p_i &= P(Y_i = 0) \\ &= \omega_i + (1-\omega_i)e^{-\lambda_i} \end{aligned} \tag{5-2}$$

and μ_i represents the mean of non-zero counts (i = 1, 2). In the non-zero part, Y_i follows the zero-truncated Poisson distribution and the mean μ_i and the variance σ_i^2 are given by

$$\mu_i = E[Y_i | Y_i > 0] = \frac{\lambda_i}{1-e^{-\lambda_i}}, \tag{5-3}$$

$$\sigma_i^2 = V[Y_i | Y_i > 0] = \frac{\lambda_i}{1-e^{-\lambda_i}} - \left(\frac{\lambda_i}{1-e^{-\lambda_i}}\right)^2 e^{-\lambda_i}. \tag{5-4}$$

The test statistic for the null hypothesis, called the two-part statistic, is defined as

$$X_{(2)}^2 = X_{(1)}^2 + U^2.$$

Here, $X^2_{(1)}$ is the test statistic for the zero part and U^2 is the test statistic for the non-zero part. The test statistic for the zero part $X^2_{(1)}$ is

$$X_{(1)}^2 = \frac{(p_1 - p_2)^2}{\frac{\overline{p}(1-\overline{p})}{N}\left(\frac{1}{r} + \frac{1}{1-r}\right)} \tag{5-5}$$

where $\overline{p} = rp_1 + (1-r)p_2$. The test statistic for the non-zero part U^2 is derived from the square of the Wilcoxon test statistic or the t-test statistic. Assuming $X^2_{(1)}$ and U^2 to be mutually independent, $X^2_{(2)}$ follows the chi-square distribution with two degrees of freedom under the null hypothesis.

The sample sizes of the non-zero portions are

$$n_1' = Nr(1 - p_1),$$
$$n_2' = N(1 - r)(1 - p_2).$$

Given two independent random samples Y_{1j} and Y_{2k} ($j = 1,\ldots, n_1'$; $k = 1,\ldots, n_2'$), the Wilcoxon test statistic for the non-zero part U is

$$U = \frac{\#(Y_1 > Y_2) + 0.5 \cdot \#(Y_1 = Y_2) - 0.5 n_1' n_2'}{\sqrt{\frac{n_1' n_2' (n_1' + n_2' + 1)}{12}\left[N'^3 - N' - \sum_t (t^3 - t) \right]}}$$

where $\#(Y_1 > Y_2)$ and $\#(Y_1 = Y_2)$ denote the number of pairs of $Y_{1j} > Y_{2k}$ and that of $Y_{1j} = Y_{2k}$, respectively, and t denotes the number of ties for each pair of $Y_{1j} = Y_{2k}$.

When calculating the sample size, generally tie data are not taken into account due to the difficulty in the assumption for ties. However, if the response is an ordered categorical variable or λ_i is small, there are considerable ties after rank transformation. In that case, the sample size may be underestimated. Zhao et al. (2008) suggested that the sample size of the Wilcoxon test be adjusted for ties. We perform this adjustment on the sample size for the two-part statistic using U' instead of U for calculating the power:

$$U' = \frac{A - 0.5}{\sqrt{\frac{1 - \sum_{c=1}^{D} \left(r P_t(c|\lambda_1) + (1-r) P_t(c|\lambda_2) \right)^3}{12 N r (1-r) p'}}} \tag{5-6}$$

where

$$A = \sum_{c=2}^{D} P_t(c|\lambda_1) \sum_{d=1}^{c-1} P_t(d|\lambda_2) + 0.5 \sum_{c=1}^{D} P_t(c|\lambda_1) P_t(c|\lambda_2).$$

Here, A denotes $\Pr(Y_1 > Y_2) + 0.5 \Pr(Y_1 = Y_2)$, and A becomes 0.5 under the null hypothesis. The denominator in the right side of equation (5-6) makes U' follow the standard normal distribution. In equation (5-6), $P_t(y|\lambda_i)$ denotes the probability density of the zero-truncated distribution

$$P_t(Y = y|\lambda_i, Y > 0) = \frac{P(y)}{1 - P(0)} = \frac{\lambda_i^y}{y!(e^{\lambda_i} - 1)}. \tag{5-7}$$

Theoretically D is infinity; however, a sufficiently large value can be used for D, while the sufficient value depends on λ_i. Generally, the λ parameter of the ZIP distribution is not so large because, even though the ZIP distribution is a bimodal distribution of the zero and non-zero parts, it is natural that the bimodal peaks are not very far apart. When $\lambda_i \leq 10$, $D = 20$ is sufficient to calculate the sample size, since the value of $P_t(D > 20)$ is ignorable (<0.001) for the sample size estimation.

The power for the two-part statistic is given by

$$Power = P_{Chi}\left(\chi^2_{\alpha, 2df}, 2, \mathrm{X}^2_{(2)} \right) \tag{5-8}$$

where $\chi^2_{\alpha,2df}$ denotes the α percentage point of the chi-square distribution with two degrees of freedom. The notation P_{Chi} indicates the cumulative probability of the chi-square distribution at $\chi^2_{\alpha,2df}$. The second parameter of two is the degrees of freedom. The third parameter of $X^2_{(2)}$ is the non-centrality parameter, which can be calculated as

$$\begin{aligned} X^2_{(2)} &= X^2_{(1)} + U'^2 \\ &= Nr(1-r)\left\{ \frac{(p_1-p_2)^2}{\bar{p}(1-\bar{p})} + \frac{12p'(A-0.5)^2}{1-\sum_{c=1}^{D}\left(rP_t(c|\lambda_1)+(1-r)P_t(c|\lambda_2)\right)^3} \right\}. \end{aligned} \tag{5-9}$$

from equations (5-5) and (5-6), where $p' = (1-p_1)(1-p_2)(1-\bar{p})^{-1}$. At the planning stage, the sample size N is determined so that the power by equation (5-8) meets the requirements of the study.

Lachenbruch (2001a) also provided the two-part statistic using the *t*-test. When the two-part statistic uses the *t*-test statistic T, it is given by

$$\begin{aligned} X^2_{(2T)} &= X^2_{(1)} + T^2 \\ &= Nr(1-r)\left[\frac{(p_1-p_2)^2}{\bar{p}(1-\bar{p})} + \frac{(\mu_1-\mu_2)^2}{\bar{\sigma}^2}p' \right]. \end{aligned} \tag{5-10}$$

In clinical research, the *t*-test is sometimes used for data following the Poisson distribution. However, since the λ parameter of the ZIP distribution is generally not so large, the normal approximation of the Poisson distribution may not be accurate in this case. The *t*-test should thus be used with care for the two-part statistic.

5.3 Comparison

5.3.1 Comparison of two-part statistics using Wilcoxon test, Wilcoxon test adjusted for ties, and *t*-test

In Section 5.2, the power of two-part statistic using the Wilcoxon test adjusted for ties was introduced. We compared the power with and without the adjustment for ties. For calculating the power of the Wilcoxon test without the adjustment for ties, the following U'' (Noether, 1987) is used instead of U' in equation (5-9):

$$U'' = \frac{\left(\Pr(Y_1 > Y_2) - 0.5\right)n_1'n_2'}{\sqrt{\frac{n_1'n_2'(n_1'+n_2'+1)}{12}}}. \tag{5-11}$$

Furthermore, the power of the two-part statistic using the *t*-test based on equation (5-10) was compared with the power of the two-part statistic using the Wilcoxon test.

Figure 5-1 displays the power of the two-part statistic using the Wilcoxon test, the Wilcoxon test adjusted for ties, and the *t*-test. The power of the three methods is similar in most cases (Figure 5-1(a)). In some cases, the power of the two-part statistic using the *t*-test is the highest (Figure 5-1(b)). However, large differences in the power between the *t*-test and the other tests were not found for large λ. As mentioned previously, since the λ parameter of the ZIP distribution is generally not so large, the *t*-test should be used with caution for the two-part statistic. The power of the two-part statistic using the Wilcoxon test is slightly higher than the power of that using the Wilcoxon test adjusted for ties. The power of the two-part statistic using the adjusted Wilcoxon test is very similar to the actual power based on a simulation study when the two-part statistic employs the Wilcoxon test. Therefore, the sample size can be reliably calculated using equation (5-6) whenever ties need to be taken into account.

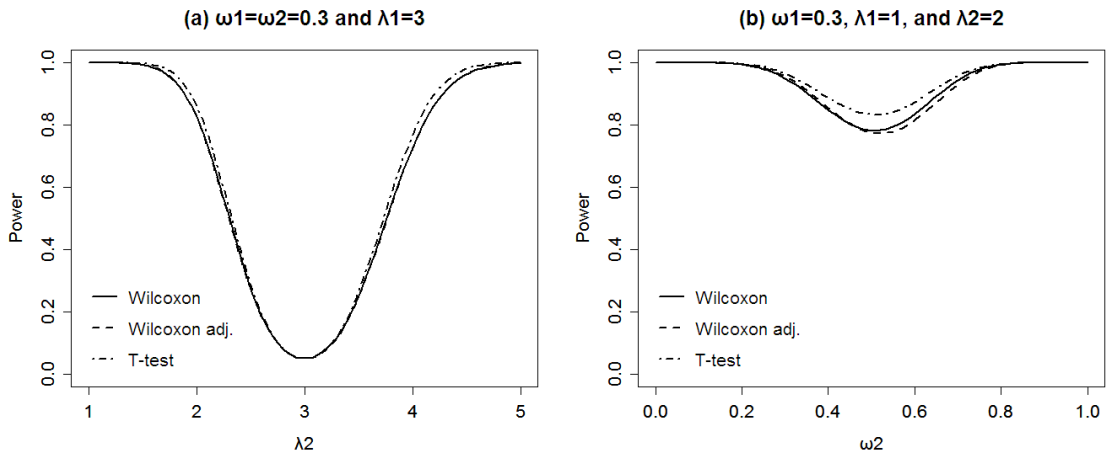


Figure 5-1. Power of the two-part statistic using the Wilcoxon test, Wilcoxon test adjusted for ties, and *t*-test when $N = 200$, $r = 0.5$, and $\alpha = 0.05$

5.3.2 Comparison with conventional tests

We compared the power of the two-part statistic with that of conventional methods ignoring zero-inflation in order to characterize the power of the two-part statistic. One of the conventional methods was the chi-square test postulating the null hypothesis of $H_{0C}: p_1 = p_2$. In this comparison, the power for the chi-square test was calculated using

the normal approximation method. Another conventional method was the Wilcoxon test for the null hypothesis of H_{0W}: $\mu_{a1} = \mu_{a2}$, where μ_{ai} was the mean of all data following the ZIP distribution:

$$\mu_{ai} = (1-\omega_i)\lambda_i . \tag{5-12}$$

The variance of the ZIP distribution is

$$\sigma_{ai}^2 = (1-\omega_i)\lambda_i(1+\omega_i\lambda_i).$$

In this case, the power of the Wilcoxon test without ties estimated by equation (5-11) is considerably different from the actual power. This difference in the power is much greater than that for the ordinary Poisson data without zero-inflation. The reason is that the ZIP data includes far more ties than the ordinary Poisson data due to many zero counts. Therefore, methods adjusted for ties should be applied to zero-inflated data whenever the sample size or power is calculated. In this article, the power was calculated using the method proposed by Zhao et al. (2008).

Figure 5-2 compares the two-part statistic using the Wilcoxon test adjusted for ties with the chi-square test and the Wilcoxon test when N = 200, r = 0.5, and α = 0.05. When $\omega_1 = \omega_2$ and $\lambda_1 = \lambda_2$ (λ_2 = 3 in Figure 5-2(a) and ω_2 = 0.3 in Figure 5-2(c)), the null hypotheses of H_0, H_{0C}, and H_{0W} are true and the power is 0.05. In Figure 5-2(d), H_{0C} is true for ω_2 = 0.322 from equation (5-1), and H_{0W} is true for ω_2 = 0.475 from equation (5-12).

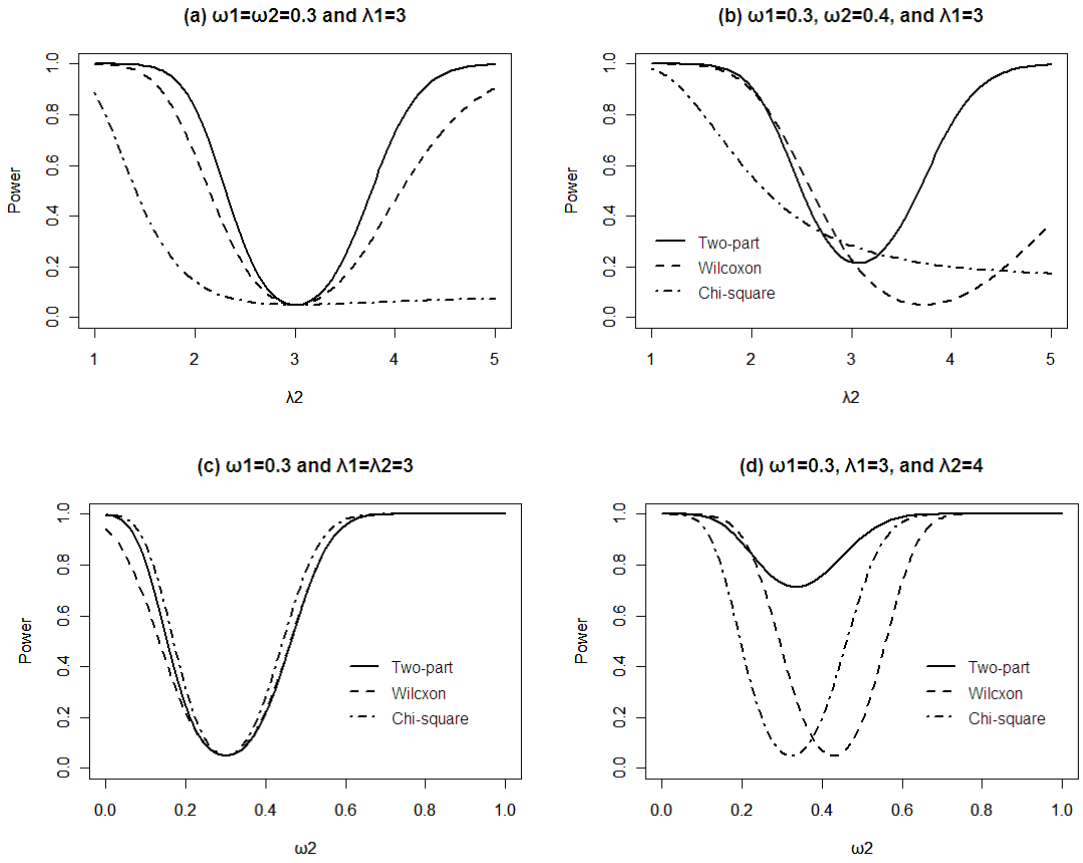


Figure 5-2. Power of the two-part statistic and conventional methods of chi-square test and Wilcoxon test when $N = 200$, $r = 0.5$, and $\alpha = 0.05$

Given no difference in ω, the two-part statistic had the highest power (Figure 5-2(a)). In the case of no difference in λ, the chi-square test gave the highest power (Figure 5-2(b)). Thus, the chi-square test was effective if the drug effect in the zero part was certain but the drug effect in the non-zero part was uncertain, although the two-part statistic was not significantly inferior to the chi-square test. Given a difference in both of ω and λ, the Wilcoxon test or the two-part statistic had the highest power. Overall, the two-part statistic maintained a steady power.

Figure 5-3(a) shows the power of the two-part statistic for various values of N. The behavior of the power did not change depending on N for different values of ω and λ. Figure 5-3(b) shows the power for the two-part statistic, the chi-square test, and the Wilcoxon test when N is small. The relationship between the three methods was the same as that presented in Figure 5-2(a).

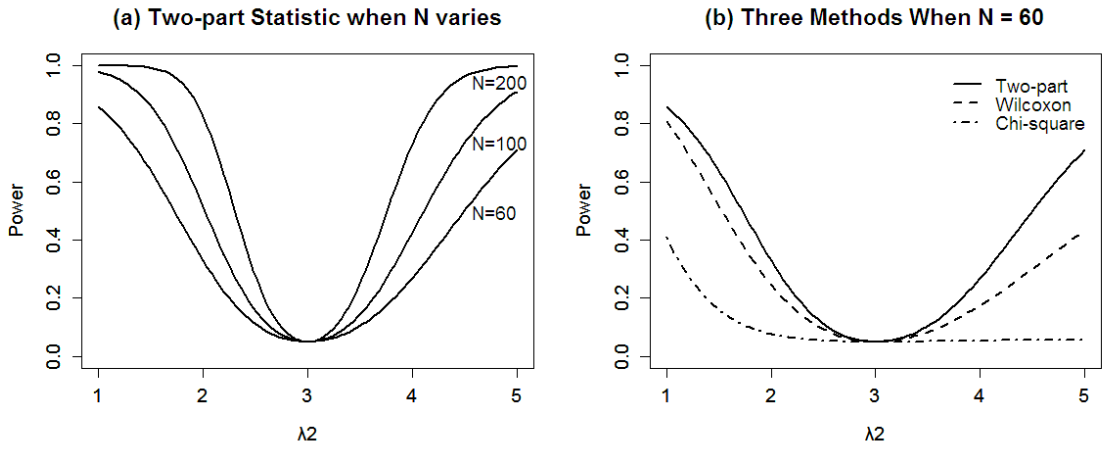


Figure 5-3. Power of the two-part statistic and conventional methods of chi-square test and Wilcoxon test when $r = 0.5$, $\alpha = 0.05$, $\omega_1 = \omega_2 = 0.3$ and $\lambda_1 = 3$.

5.3.3 Comparison with the ZIP model

In this section, the ZIP model for assessing the treatment effect is described, and a comparison of the two-part statistic and the ZIP model is made using a simulation study. The ZIP model is a mixture model of the logit model for ω and the Poisson regression model for λ. The ZIP model is expressed by

$$\text{logit}(\omega) = \mathbf{G}\boldsymbol{\gamma},$$
$$\log(\lambda) = \mathbf{B}\boldsymbol{\beta},$$

where **B** and **G** represent covariate matrices. Lambert (1992) supposed that one could observe $Z_l = 1$ when Y_l was from the inflated zero, and $Z_l = 0$ when Y_l was from the Poisson state. The parameters γ and β are estimated by maximizing the log-likelihood:

$$L(\boldsymbol{\gamma},\boldsymbol{\beta};y,z) = \sum_{l=1}^{N}\log\left(f\left(z_l|\boldsymbol{\gamma}\right)\right) + \sum_{l=1}^{N}\log\left(f\left(y_l|z_l,\boldsymbol{\beta}\right)\right)$$
$$= \sum_{l=1}^{N}\left[z_l\mathbf{G}_l\boldsymbol{\gamma} - \log\left(1+e^{\mathbf{G}_l\boldsymbol{\gamma}}\right)\right] + \sum_{l=1}^{N}(1-z_l)\log\left(y_l\mathbf{B}_l\boldsymbol{\beta} - e^{\mathbf{B}_l\boldsymbol{\beta}}\right) - \sum_{l=1}^{N}(1-z_l)\log y_l!$$

where $\mathbf{G}_l$ and $\mathbf{B}_l$ are the l-th rows of **G** and **B**. On the right side of the equation, the first term including γ and the second term including β can be maximized separately. Recently, maximum likelihood estimation procedures have been included in software packages such as the package pscl in R (Zeileis et al., 2008) and the COUNTREG procedure in SAS.

The treatment effect T can be included in the covariate matrices **G** and **B**. The null

hypotheses for the treatment comparison are $H_{0\gamma}$: $\gamma_t = 0$ and $H_{0\beta}$: $\beta_t = 0$, where γ_t and β_t are the parameter estimates of the treatment effect. That is, there are two hypothesis tests for the treatment effect: one in zero-inflation and one in the mean of the Poisson distribution. On the other hand, the two-part statistic is one comprehensive test. In order to compare the power of testing the treatment effect in the ZIP model with the power of the two-part statistic, Fisher's combination test (Fisher, 1948) is applied. Assume that the p-values p_γ for $H_{0\gamma}$ and p_β for $H_{0\beta}$ are mutually independent and uniformly distributed on [0, 1] under the null hypothesis. The level α combination test rejects the null hypotheses $H_{0\gamma}$ and $H_{0\beta}$ if

$$p_\gamma p_\beta < \exp\left(-\frac{1}{2}\chi^2_{1-\alpha,4}\right). \tag{5-13}$$

We compared power using a simulation study where the power was defined as rejected proportions of the null hypothesis on 10,000 replications. The estimates of the ZIP model were computed using the *zeroinfl* function of the package pscl in R. The ZIP model included the treatment effect alone. The two-part statistic was calculated using the chi-square test and the Wilcoxon test. The allocation ratio r was 0.5, and α was set at 0.05.

Table 5-1 shows the simulation results for $N = 200$. The table contains the results for three cases: treatment difference in the zero-inflation ω alone ($\lambda_1 = \lambda_2 = 3$), treatment difference in the Poisson parameter λ alone ($\omega_1 = \omega_2 = 0.3$), and treatment difference in both parameters ($\omega_1 = 0.3$, $\omega_2 = 0.1 - 0.5$, $\lambda_1 = 3$, and $\lambda_2 = 1 - 3$). When $N = 200$, the power of the combination test of the ZIP model and of the two-part statistic were generally similar. Differences were found between the test for zero-inflation of the ZIP model and the chi-square test component of the two-part statistic. The difference was large when ω_1 and ω_2 were equal or similar and λ_2 was 1 or 2. This was caused by the difference in the null hypotheses of $H_{0\gamma}$: $\gamma_t = 0$ and H_0: $p_1 = p_2$, where p_1 and p_2 included the zero counts from the Poisson state. If the treatment difference in the zero-inflation is of particular interest, then the ZIP model can detect this.

Table 5-1. Power of the zero-inflated Poisson model and the two-part statistic when a covariate was treatment effect alone, $N = 200$, $r = 0.5$, and $\alpha = 0.05$

				Power for Two-part Statistic			Power for ZIP-model		
ω_1	ω_2	λ_1	λ_2	Two-part	χ^2	Wilcoxon	Combination	Zero-inflation	Non-zero
0.3	0.3	3	1	>0.999	0.899	>0.999	>0.999	0.040	>0.999
0.3	0.3	3	2	0.835	0.140	0.877	0.820	0.038	0.895
0.3	0.3	3	3	0.047	0.051	0.047	0.046	0.045	0.052
0.3	0.3	3	4	0.737	0.065	0.816	0.751	0.046	0.847
0.3	0.3	3	5	0.999	0.076	>0.999	>0.999	0.045	>0.999
0.3	0.4	3	1	>0.999	0.989	>0.999	>0.999	0.158	>0.999
0.3	0.4	3	2	0.926	0.564	0.851	0.894	0.243	0.878
0.3	0.4	3	3	0.214	0.281	0.048	0.199	0.272	0.048
0.3	0.4	3	4	0.760	0.196	0.786	0.806	0.283	0.810
0.3	0.4	3	5	>0.999	0.166	0.999	0.999	0.291	>0.999
0.3	0.1	3	3	0.830	0.898	0.048	0.685	0.838	0.049
0.3	0.2	3	3	0.234	0.315	0.043	0.204	0.289	0.050
0.3	0.3	3	3	0.047	0.051	0.047	0.046	0.045	0.052
0.3	0.4	3	3	0.213	0.281	0.048	0.199	0.272	0.048
0.3	0.5	3	3	0.674	0.778	0.049	0.654	0.761	0.049
0.3	0.1	3	4	0.995	0.969	0.862	0.992	0.907	0.885
0.3	0.2	3	4	0.894	0.474	0.846	0.889	0.322	0.862
0.3	0.3	3	4	0.737	0.065	0.816	0.751	0.046	0.847
0.3	0.4	3	4	0.760	0.196	0.786	0.806	0.283	0.810
0.3	0.5	3	4	0.914	0.710	0.748	0.936	0.784	0.774

Table 5-2 shows the simulation results for $N = 60$. The two-part statistic had higher power than the combination test for the ZIP model in some cases (e.g., $\omega_1 = \omega_2 = 0.3$, $\lambda_1 = 3$, and $\lambda_2 = 1$; $\omega_1 = 0.3$, $\omega_2 = 0.4$, $\lambda_1 = 3$, and $\lambda_2 = 1$). This magnitude of difference was not found for $N = 200$, because the power for the two-part statistic and the combination test for the ZIP model was in excess of 0.99 in these cases. Regarding the combination test for the ZIP model, the power under the null hypothesis (i.e., $\omega_1 = \omega_2 = 0.3$, $\lambda_1 = \lambda_2 = 3$) was less than the nominal type I error rate of 0.05 although this was approximately 0.05 for $N = 200$. In the case of no treatment difference in the zero-inflation scenario (i.e., $\omega_1 = \omega_2$), the power for the zero-inflation of the ZIP model was also less than the nominal type I error rate of 0.05 in Table 5-2. In contrast, regarding the two-part statistics, the power under the null hypothesis was approximately

0.05 and did not change for N= 200.

Table 5-2. Power of the zero-inflated Poisson model and the two-part statistic when a covariate was treatment effect alone, N = 60, r = 0.5, and α = 0.05

				Power for Two-part Statistic			Power for ZIP-model		
ω_1	ω_2	λ_1	λ_2	Two-part	χ^2	Wilcoxon	Combination	Zero-inflation	Non-zero
0.3	0.3	3	1	0.932	0.417	0.910	0.857	0.032	0.929
0.3	0.3	3	2	0.316	0.071	0.382	0.278	0.020	0.412
0.3	0.3	3	3	0.047	0.046	0.046	0.032	0.027	0.044
0.3	0.3	3	4	0.250	0.049	0.325	0.255	0.032	0.357
0.3	0.3	3	5	0.736	0.054	0.826	0.773	0.037	0.863
0.3	0.4	3	1	0.948	0.611	0.874	0.837	0.079	0.888
0.3	0.4	3	2	0.390	0.206	0.344	0.291	0.082	0.368
0.3	0.4	3	3	0.094	0.117	0.045	0.070	0.093	0.042
0.3	0.4	3	4	0.275	0.093	0.315	0.307	0.090	0.345
0.3	0.4	3	5	0.720	0.087	0.801	0.777	0.097	0.849
0.3	0.1	3	3	0.320	0.408	0.046	0.065	0.941	0.044
0.3	0.2	3	3	0.098	0.125	0.045	0.047	0.052	0.046
0.3	0.3	3	3	0.047	0.046	0.046	0.032	0.027	0.044
0.3	0.4	3	3	0.094	0.117	0.045	0.070	0.093	0.042
0.3	0.5	3	3	0.245	0.320	0.046	0.191	0.280	0.046
0.3	0.1	3	4	0.649	0.535	0.363	0.476	0.234	0.377
0.3	0.2	3	4	0.372	0.176	0.344	0.333	0.095	0.371
0.3	0.3	3	4	0.250	0.049	0.325	0.255	0.032	0.357
0.3	0.4	3	4	0.275	0.093	0.315	0.307	0.090	0.345
0.3	0.5	3	4	0.405	0.278	0.283	0.442	0.286	0.314

5.4 Example

We consider a clinical trial whose aim is to compare a hormone therapy to a placebo for menopausal women with vasomotor symptoms. Hot flushes are common in menopausal women (CHMP, 2005). The daily frequency of moderate to severe hot flushes is the primary endpoint of the trial. When the trial subjects are patients with mild symptoms, in the range of three to twelve hot flushes daily, hot flushes can be completely removed in a certain proportion of subjects by hormone therapy and even by

placebo treatment.

Assume that the theoretical data in Figure 5-4 have been obtained in a previous study. The ZIP parameters estimated from the data are $\omega_1 = 0.42$ and $\lambda_1 = 3.77$ for the hormone therapy and $\omega_2 = 0.21$ and $\lambda_2 = 4.57$ for the placebo using the zeroinfl function in R or the COUNTREG procedure in SAS.

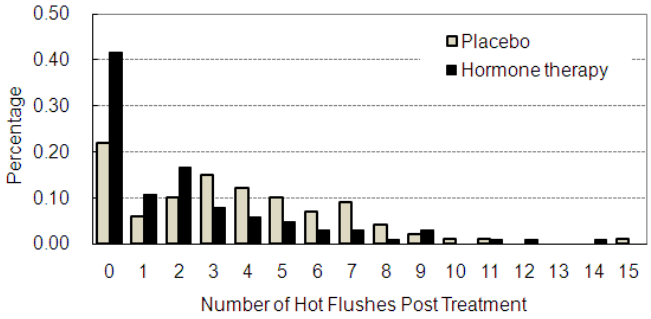


Figure 5-4. Daily frequency of moderate to severe hot flushes

If individual data are not available but the mean of non-zero counts μ_i and the proportion of zero counts p_i are known, the ZIP parameters can be roughly estimated as follows. Suppose that we know only $\mu_1 = 3.86$ and $p_1 = 0.43$ for the hormone therapy, and $\mu_2 = 4.62$ and $p_2 = 0.22$ for the placebo. The mean μ_i is equal to $\lambda_i/[1 - \exp(-\lambda_i)]$ by equation (5-3). Using the iteration scheme, $\lambda_i^{(t+1)} = \mu_i [1 - \exp(\lambda_i^{(t)})]$, where $\lambda_i^{(t)}$ is the t-th estimate (Dietz et al., 2000), λ_1 converges to 3.77 for the hormone therapy and λ_2 to 4.57 for the placebo. Then, $\omega_1 = 0.42$ for the hormone therapy and $\omega_2 = 0.20$ for the placebo are estimated from $\omega_i = [p_i - \exp(-\lambda_i)]/[1 - \exp(-\lambda_i)]$ based on equation (5-2).

The sample size for the two-part statistic is calculated from equations (5-8) and (5-9). In equation (5-6), $A = 0.601$. For a 90% power, the sample size adjusting for ties is 158 when $r = 0.5$ and $\alpha = 0.05$.

5.5 Discussion

In clinical trials, it is often undesirable to perform two hypothesis tests in the primary analysis because of the difficulty in interpreting the two possibly inconsistent test results and controlling the type I error inflation. The two-part statistic is effective in terms of producing one test result as well as embracing the treatment effect in the zero and the non-zero parts. We provided the methods for calculating sample size and power

for the two-part statistic using the Wilcoxon test adjusted for ties. The power estimated by our method was very similar to the actual power based on a simulation study when the two-part statistic employed the Wilcoxon test. In our study, the two-part statistic showed higher power than conventional tests in most cases. Even when a conventional test showed higher power than the two-part statistic, the difference was small. However, if the complete recovery from a disease is of primary interest in the clinical research, the chi-square test on a binary response of zero vs. non-zero outcome should be used for the primary analysis.

The ZIP model can estimate the profile of the ZIP distribution and the extent of zero-inflation. For example, Cheung (2002) applied the ZIP model to child growth and motor development and explored the effect of covariates. However, our study focused on the treatment effect in the ZIP model for comparison with the two-part statistic. The simulation results showed that the power of the combination test for the ZIP model and the two-part statistic were generally similar when the covariate was the treatment effect alone. In some results with a small sample size, the two-part statistic demonstrated higher power than the ZIP model. The ZIP model is suitable for characterizing the distribution and exploring the effect of some covariates but may be unsuitable for aiming at the detection of treatment difference especially when the sample size is not large. In our simulation study, we only considered the treatment effect in the ZIP model for the purpose of comparison with the two-part statistic. However, in practice, the ZIP model can include some covariates in the logit model for ω and the Poisson regression model for λ.

The two-part statistic can be adjusted with covariates. For example, a logistic regression model is used for the zero part. For the non-zero part, ANCOVA with ranks transformed from Y can be applied (Conover et al., 1982). It is a subject for further research to develop methods for the sample size estimation in this context and to assess the effect of covariates on the power.

6 CONCLUSION

The purpose of this study was to enhance the efficiency of the confirmatory clinical trials in order to overcome the difficulties in new drug development: e.g., the scale of phase 3 studies have been getting large; regulatory authorities request large safety data; the drug development is becoming increasingly expensive, and has a high clinical failure rate. For the purpose, we studied the adaptive design and statistical testing and sample size for a special distribution. This article dealt with three subjects about the adaptive design: optimal timing for interim analyses, sample size re-estimation for survival data, and clinically important effects. Regarding the optimal timing for interim analyses, we formulated ASN under practical conditions and examined the timing for interim analyses in terms of minimizing the ASN. As for the sample size re-estimation for survival data, we applied the methodology of sample size re-estimation for continuous data to survival data, and proposed an interim hazard ratio estimate that could be used to re-estimate the sample size. For clinical important effects, we stated the distinction between MCIC and MCID, and their roles in new drug development. We studied the approaches to estimate MCIC and MCID. Regarding statistical testing and sample size for a special distribution, we focused on the zero-inflated count data because zero inflation is frequently ignored when comparing treatment groups. We provided the methods for calculating sample size and power for the two-part statistic using the Wilcoxon test adjusted for ties.

The first part of this article addressed the subjects relative to adaptive design. First, we demonstrated the methods used to find the optimal time for interim analyses in order to minimize the ASN. When an interim analysis was performed in Case 1, the optimal time was approximately $t_1 = 2/3$ for the O'Brien-Fleming type and approximately $t_1 = 1/2$ for the Pocock type, regardless of the effect size. These results based on the ASN were consistent with the common impression that the optimal time for the O'Brien-Fleming type would be later than that for the Pocock type because of spending less type I error rate at the interim analysis with the O'Brien-Fleming type. When two interim analyses were performed in Case 1, the time of the second interim analysis had little effect on the ASN. When the sample size is planned based on the clinically meaningful effects, the optimal time of the interim analysis is earlier than when the sample size is planned based on the anticipated effect. The group sequential design is more efficient compared with a fixed design when the anticipated effect is much greater than the clinically meaningful effect. We showed that the optimal time for the interim

analysis depended on the follow-up duration in Case 2. The interim analysis did not considerably reduce the ASN when the follow-up duration was longer than one-third of the enrollment duration.

Second, we addressed the methodology of sample size re-estimation for survival data, and proposed an interim hazard ratio estimate that could be used to re-estimate the sample size under those circumstances. In the proposed method to estimate interim hazards, constant c is set in proportion to the degree of confidence in the hypothesized hazard and the observed data. The information fraction of the interim analysis can be one possible factor to determine the value of c. The CHW method applied to the log-rank test was comparable to other methods in terms of the overall power and control of the type I error rate. Since the CHW method uses α for the fixed design, it can easily be applied in actual clinical trials. However, this may cause critical problems from a regulatory view point if the sample size re-estimation allows the sample size to be reduced from the preplanned sample size. If the interim analysis does not result in an early termination for success but results in a modified number of events equal to zero or a very few, there may be serious doubt about leaving the final significance level as the preplanned α_2 in the CHW method. Cui *et al*. (1999) does not suppose a decrease in the sample size at the interim analysis. If the CHW method is used, the sample size should not be reduced from the preplanned one at the interim analysis.

Third, clinically important effects were described. We stated the distinction between MCIC and MCID, and their roles in new drug development. Furthermore, we presented how the clinically important effect of the drug should be demonstrated using MCIC or MCID in clinical development. We would recommend discussing the clinically important effect at the planning phase of trials. The MCIC and MCID may sometimes change because the results of PoC trials may reveal new profiles of the drug or the standard treatment may improve during the drug development. Hence, investigators and sponsors should keep reviewing the MCIC and MCID even after those are established. Although this article focused on the clinical importance of efficacy, it is also necessary to evaluate safety and the risk-benefit.

The latter part of this article presented methods of statistical testing and sample size for a special distribution focusing on the zero-inflated count data. We provided the methods for calculating sample size and power for the two-part statistic using the Wilcoxon test adjusted for ties. The power estimated by our method was very similar to the actual power based on a simulation study when the two-part statistic employed the Wilcoxon test. In our study, the two-part statistic showed higher power than

conventional tests in most cases. Even when a conventional test showed higher power than the two-part statistic, the difference was small. However, if the complete recovery from a disease is of primary interest in the clinical research, the chi-square test on a binary response of zero vs. non-zero outcome should be used for the primary analysis. The two-part statistic can be adjusted with covariates. For example, a logistic regression model is used for the zero part. For the non-zero part, ANCOVA with ranks transformed from Y can be applied (Conover et al., 1982). It is a subject for further research to develop methods for the sample size estimation in this context and to assess the effect of covariates on the power.

We hope that these findings will enhance more efficient study design in new drug development.

ACKNOWLEDGEMENT

I am grateful to Professor Manabu Iwasaki for teaching, supporting, and encouraging me during the last two decades. I greatly appreciate insightful and helpful comments given by Professor Chikuma Hamada (Tokyo University of Science), Professor Ichie Watanabe (Seikei University), and Professor Atsuko Ikegami (Seikei University).

REFERENCES

Bauer M, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50:1029-1041.

Burback D, Molnar FJ, John PS, Man-Son-Hing M. Key methodological features of randomized controlled trials of Alzheimer's disease. *Dement Geriatr Cogn Disord.* 1999;10:534-540.

Cartwright ME, Cohen S, Fleishaker JC, et al. Proof of concept: A PhRMA position paper with recommendations for best practice. *Clin Pharmacol Ther.* 2010;87(3):278-285.

Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*. 2004; 23:1023-1038.

Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*. 2004;21:1461–1469.

Chow SC, Liu JP. *Design and Analysis of Clinical Trials: Concepts and Methodoloies, Second Edition*. New Jersey: John Wiley & Sons, Inc.; 2004:438-439.

Cohen J. *Statistical Power Analysis for the Behavioral Sciences, Second Edition*. New Jersey: Lawrence Erlbaum Associates, Inc.; 1988:19-27.

Colton T, McPherson K. Two-stage plans compared with fixed-sample-size and Wald SPRT plans. *Journal of the American Statistical Association*. 1976;71(353):80-86.

Cox DR. Partial likelihood. Biometrika 1975; 62:269-276.

Committee for Medical Products for Human Use (CHMP). Guideline on clinical investigation of medical products for hormone replacement therapy of oestrogen deficiency symptoms in postmenopausal women. European Medicines Agency: London. 2005.

Conover WJ and Iman RL. Analysis of covariance using the rank transformation. *Biometrics.* 1982;38:715-724.

Cui L, Hung HMJ, Wang S. Modification of sample size in group sequential clinical trials. *Biometrics* 1999;55:853-857.

D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues — the encounters of academic consultants in statistics. *Statistics in Medicine.* 2003;22:169-186.

Delucchi KL, Bostrom A. Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *Am J Psychiatry*i. 2004;161:1159–1168.

Desseaux K, Porcher R. Flexible two-stage design with sample size reassessment for survival trials. *Statistics in Medicine*. 2007;26:5002-5013.

Dietz E, Böhning D. On estimation of the Poisson parameter in zero-inflated Poisson models. *Computational Statistics and Data Analysis*. 2000;34:441-459.

European Medicines Agency (EMA). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. EMA. 2007. (Available at: http://www.ema.europa.eu/docs/en_GB/document_library/ Scientific_guideline/2009/09/WC500003616.pdf)

Farrar JT, Young JP, LaMoreaux L, et al. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*. 2001;149-158.

Fisher RA. Questions and answers: Combining independent tests of significance. *The American Statistician*. 1948;2:30.

Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinherio J. Adaptive design in clinical drug development − An executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics*. 2006,16:275-283.

Golub HL. The need for more efficient trial designs. *Statistics in Medicine.* 2006; 25:3231-3235.

Gould AL. Timing of futility analyses for 'proof of concept' trials. *Statistics in Medicine*. 2005;24:1815-1835.

Hallstorm AP. A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine.* 2010;29:391–440.

Heilbron DC. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*. 1994;36:531-547.

Hung HMJ, Wang SJ, O'Neill RT. Methodological issues with adaptation of clinical trial design. *Pharmaceut Statist.* 2006;5:99-107.

International Conference on Harmonisation of Technical Requirements for Registratoin of Pharmaceuticals for Human Use (ICH-E8). General consideration for clinical trials (E8). ICH. 1997. (Available at: http://www.ich.org/fileadmin/ Public_Web_Site/ICH_Products/Guidelines/Efficacy/E8/Step4/E8_Guideline.pdf)

Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol.* 1991;59(1):12-19.

Jahn-Eimermacher A, Hommel G. Performance of adaptive sample size adjustment with respect to stopping criteria and time of interim analysis. *Statistics in Medicine* 2007;26:1450-1461.

Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine.* 2006;25:917-932.

Kober A, Torp-Pedersen C, *et al*. A clinical trial of the angiotensin-converting-enzyme inhibitor trandolapril in patients with left ventricular dysfunction after myocardial infarction. *The New England Journal of Medicine*. 1995; 333(25):1670-1676.

Lachenbruch PA. Comparisons of two-part models with competitors. *Statistics in Medicine*. 2001a;20:1215–1234.

Lachenbruch PA. Power and sample size requirements for two-part models. *Statistics in Medicine* 2001b;20:1235–1238.

Lachin, JM, Foulkes, MA. Evaluation of sample size and power for analysis of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics.* 1986;42:507–519.

Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics.* 1992;34:1-14.

Lan KKG., DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika.* 1983;70(3):659-663.

Lawrence J. Strategies for changing the test statistic during a clinical trial. *Journal of Biopharmaceutical Statistics.* 2002;12(2):193-205.

Li G, Shih WJ, Wang Y. Two-stage adaptive design for clinical trials with survival data. *Journal of Biopharmaceutical Statistics.* 2005;15:707-718.

Li G, Shih WJ, Xie T, Lu J. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics.* 2002;3(2):277-287.

Make B. How can we assess outcomes of clinical trials: The MCID approach. *Journal of Chronic Obstructive Pulmonary Disease.* 2007;4:191-194.

Man-Son-Hing M, Laupacis A, O'Rourke K, *et al*. Determination of the clinical importance of study results. *J Gen Intern Med.* 2002;17:469-476.

Noether GE. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*. 1987;82(398):645–647.

Pitt B, Remme W, Zannad F, *et al*. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infraction. *The New England Journal of Medicine.* 2003;348(14):1309-1321.

Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics.* 1995;51:1315-1324.

Proschan MA, Lan KKG, Wittes JT. *Statistical Monitoring of Clinical Trials*. Springer; 2006.

Schäfer H, Müller H-H. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine.* 2001;20:3741-3751.

Shen Y, Cai J. Sample size reestimation for clinical trials with censored survival data. *Journal of the American Statistical Association.* 2003;98:418-426.

Shih WJ. Sample size re-estimation ― journey for a decade. *Statistics in Medicine.* 2001;20:515-518.

Shih WJ. Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison. *Statistics in Medicin.e* 2006;25:933-941.

Sierevelt IN, Oldenrijk J, Poolman RW. Is statistical significance clinically important? ― A guide to judge the clinical relevance of study findings. *Journal of Long-Term Effects of Medical Implants.* 2007;17(2):173-179.

Togo K, Iwasaki M. Sample size re-estimation for survival data in clinical trials with an adaptive design. Pharmaceutical Statistics. 2011;10(4):325-31.

Togo K, Matsuoka N, Hashigaki S, Imai K, Moriya T. Clinically Important Effects in New Drug Development. Drug Information Journal. 2011,45:805-10.

Togo K, Iwasaki M. Optimal Timing for Interim Analyses in Clinical Trials. Journal of Biopharmaceutical Statistics. 2013;23:1067-80.

Togo K, Iwasaki M. Group Comparisons Involving Zero-inflated Count Data in Clinical Trials. Japanese Journal of Biometrics. 2013 (in press).

Tsiatis AA. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika.* 1981;68(1):311-315.

Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika.* 2003;90(2):367-378.

Whitehead J, Whitehead A, Todd S, Bolland K, Sooriyarachchi MR. Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine.* 2001;20:165-176.

Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *Journal of Statistical Software.* 2008;27(8).

Zhao YD, Rahardja D, Qu1 Y. Sample size calculation for the Wilcoxon–Mann–Whitney test adjusting for ties. *Statistics in Medicine.* 2008;27:462–468.