Journal of Biopharmaceutical Statistics, 16: 151–164, 2006

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400500508754

Taylor & Francis
Taylor & Francis Group

A RATIO TEST IN ACTIVE CONTROL NON-INFERIORITY TRIALS WITH A TIME-TO-EVENT ENDPOINT

Yong-Cheng Wang
Biostatistics, Centocor, Inc., Malvern, Pennsylvania, USA

Gang Chen and George Y.H. Chi
Pharmaceutical Research and Development, Johnson & Johnson, Raritan, New Jersey, USA

There are essentially two kinds of non-inferiority hypotheses in an active control trial: fixed margin and ratio hypotheses. In a fixed margin hypothesis, the margin is a prespecified constant and the hypothesis is defined in terms of a single parameter that represents the effect of the active treatment relative to the control. The statistical inference for a fixed margin hypothesis is straightforward. The outstanding issue for a fixed margin non-inferiority hypothesis is how to select the margin, a task that may not be as simple as it appears. The selection of a fixed non-inferiority margin has been discussed in a few articles (Chi et al., 2003; Hung et al., 2003; Ng, 1993). In a ratio hypothesis, the control effect is also considered as an unknown parameter, and the non-inferiority hypothesis is then formulated as a ratio in terms of these two parameters, the treatment effect and the control effect. This type of non-inferiority hypothesis has also been called the fraction retention hypothesis because the ratio hypothesis can be interpreted as a retention of certain fraction of the control effect. Rothmann et al. (2003) formulated a ratio non-inferiority hypothesis in terms of log hazards in the time-to-event setting. To circumvent the complexity of having to deal with a ratio test statistic, the ratio hypothesis was linearized to an equivalent hypothesis under the assumption that the control effect is positive. An associated test statistic for this linearized hypothesis was developed. However, there are three important issues that are not addressed by this method. First, the retention fraction being defined in terms of log hazard is difficult to interpret. Second, in order to linearize the ratio hypothesis, Rothmann's method has to assume that the true control effect is positive. Third, the test statistic is not powerful and thus requires a huge sample size, which renders the method impractical. In this paper, a ratio hypothesis is defined directly in terms of the hazard. A natural ratio test statistic can be defined and is shown to have the desired asymptotic normality. The demand on sample size is much reduced. In most commonly encountered situations, the sample size required is less than half of those needed by either the fixed margin approach or Rothmann's method.

Key Words: Active control; Fixed margin; Fraction retention; Linearized hypothesis; Non-inferiority; Ratio hypothesis; Ratio test; Superiority; Time-to-event endpoint.

Received March 1, 2005; Accepted September 1, 2005
Address correspondence to Yong-Cheng Wang, Biostatistics, Centocor, Inc., Mail Stop C-4-2, 200 Great Valley Parkway, Malvern, PA 19355, USA; E-mail: YCWang_US@yahoo.com

1. INTRODUCTION

The purpose of clinical trials in drug development is usually to demonstrate that a new treatment is superior to placebo or standard of care. The hypotheses in such clinical trials involve only one parameter of interest, e.g., treatment effect (relative to placebo or standard care). For mortality or serious morbidity trials, ethical reason requires that when there are available known effective treatments for the disease, investigators should use one of these active treatments instead of a placebo as the control (Declaration of Helsinki, 2000). Traditionally, in an active control trial, the objective is to demonstrate that the new experimental treatment is superior to the active control. However, to require that the new experimental treatment be superior to control may not always be necessary as when, for example the experimental treatment offers other advantage over the control, such as better safety profile, or ease of administration (Temple, 1996; Temple and Ellenberg, 2000).

Generally, two types of hypotheses can be formulated in an active control non-inferiority trial: a fixed margin hypothesis and a fraction retention hypothesis. With a fixed margin hypothesis, the margin is specified as a constant. The statistical inference has been well developed for such a one-parameter fixed margin hypothesis. The real issue with the fixed margin non-inferiority hypothesis is how to specify the non-inferiority margin. The selection of a fixed non-inferiority margin has been discussed in a few articles (Chi et al., 2003; Hung et al., 2003; Ng, 1993). With a fraction retention hypothesis, the control effect is considered as an unknown parameter. The hypothesis can be formulated as a ratio of two parameters, the treatment effect and the control effect. The statistical tests for a fraction retention hypothesis has recently been developed. For example, Rothmann et al. (2003) formulated a ratio non-inferiority hypothesis and developed an associated statistical test for the linearized hypothesis that is equivalent to the original ratio hypothesis under the assumption that the true control effect is positive. Rothmann's method considers the problem within the context of a time-to-event endpoint, and defines the concept of fraction retention in terms of log-hazard ratio. The major issues with Rothmann's method will be discussed later in this section.

A fraction retention non-inferiority hypothesis generally involves two parameters, a treatment effect θ_t and a control effect θ_c. The following ratio hypothesis is a direct formulation of the non-inferiority fraction retention hypothesis:

$$H_0 : \theta_t/\theta_c \geq 1 - \delta_0 \quad \text{vs.} \quad H_a : \theta_t/\theta_c < 1 - \delta_0, \tag{1}$$

where δ_0 ($0 \leq \delta_0 \leq 1$) denotes a given level of fraction retention.

The treatment effect θ_t and the control effect θ_c have different expressions for different endpoints. For instance, for a mortality trial with a time-to-event endpoint, θ_t and θ_c are usually expressed as hazard ratios. For a binary endpoint, θ_t and θ_c may represent the odds ratio or relative risk of treatment relative to control, and control relative to placebo, respectively. Let T, C, and P be the experimental treatment, control, and placebo, respectively. Rothmann et al. (2003) used $\log HR(T/C)$ (i.e., θ_t) and $\log HR(P/C)$ (i.e., θ_c) to express the treatment effect relative to the active control and the control effect relative to placebo, respectively. Although it is difficult to interpret clinically, it avoids mathematics difficulties in the

statistical inference. In this paper, we also consider a mortality trial with a time-to-event endpoint and use $HR(T/C) - 1$ (i.e., θ_t) and $HR(P/C) - 1$ (i.e., θ_c) to express the treatment effect relative to the active control and the control effect relative to placebo, respectively, which are clinically more interpretable.

In hypothesis (1), since it is difficult to draw statistical inference directly on such ratio hypothesis, a transformation of the ratio hypothesis is often considered. If we assume that $\theta_c > 0$, the transformation is straightforward. The non-inferiority hypotheses (1) can be transformed into:

$$H_0 : \theta_t - (1 - \delta_0)\theta_c \geq 0 \quad \text{vs.} \quad H_a : \theta_t - (1 - \delta_0)\theta_c < 0. \tag{2}$$

For the non-inferiority hypothesis in (2), the following test statistic can be used.

$$Z_R^* = \frac{\hat{\theta}_t - (1 - \delta_0)\hat{\theta}_c}{\sqrt{SE^2(\hat{\theta}_t) + (1 - \delta_0)^2 SE^2(\hat{\theta}_c)}}, \tag{3}$$

where $\hat{\theta}_t$ and $\hat{\theta}_c$ are the estimates of θ_t and θ_c, respectively, and $SE(\hat{\theta}_t)$ and $SE(\hat{\theta}_c)$ are the standard errors of $\hat{\theta}_t$ and $\hat{\theta}_c$, respectively. Since $\hat{\theta}_t$ and $\hat{\theta}_c$ are based on data collected from the concurrent trial and historical trial(s), respectively, they are independent and Z_R^* is asymptotically normally distributed.

For a time-to-event endpoint, Rothmann et al. (2003) define the fraction retention of control effect to be retained by the experimental treatment as

$$\delta_R = \frac{\log HR(P/C) - \log HR(T/C)}{\log HR(P/C)}, \tag{4}$$

where T, C, and P represent experimental treatment, control, and placebo, respectively, and $\log HR(P/C)$ is assumed to be positive. Furthermore, it is assumed that the constancy assumption holds, that is, the control effect has not changed over time. This assumption may not necessarily hold true in general due to changing standard of care or medical practice, such as the standard therapy for treating patient with the disease under consideration. However, all methods would implicitly or explicitly require such assumptions, unless data are available to provide appropriate adjustment.

Based on definition (4), the non-inferiority hypothesis of interest can be defined as

$$H_0 : \delta_R \leq \delta_0 \quad \text{vs.} \quad H_a : \delta_R > \delta_0, \tag{5}$$

where δ_0 is the level of fraction retention desired. Though δ_0 is defined as the level of fraction retention in many papers, δ_0 also depends on the definition of δ_R. For instance, δ_0 in non-inferiority ratio hypothesis and δ_0 in Rothmann's non-inferiority linearized hypothesis are different as shown in Chi et al. (2003).

Under the assumption that the control effect is positive, the hypothesis in (5) is equivalent to

$$H_0 : \log HR(T/C) - (1 - \delta_0)\log HR(P/C) \geq 0$$
$$\text{vs.} \quad H_a : \log HR(T/C) - (1 - \delta_0)\log HR(P/C) < 0.$$

Rothmann et al. (2003) developed a statistical test for the above hypothesis, which is analogous to Z_R^*, except the estimates of the hazard ratios are replaced by the estimates of the log hazard ratios. A similar test is also proposed for relative risk by Holmgren (1999) and Hasselblad and Kong (2001).

The linearized non-inferiority hypothesis based on the fraction retention concept greatly reduces the mathematical difficulties encountered in deriving the distributional properties of the test statistic associated with the ratio non-inferiority hypothesis. There are three major issues with this hypothesis: 1) The ratio is defined based on a log transformation of a hazard ratio. Clinically, it is difficult to interpret. 2) The hypothesis requires a key assumption that active control is truly effective. This assumption leads to the inflation of false positive rate of the trial (Chen et al., 2004). 3) The statistical test is not powerful and imposes a large sample size requirement.

In the following section, fraction retention will be defined directly in terms of the hazard ratio and an associated ratio test statistic will be defined. The proof of the asymptotic normality of this ratio test statistic will be given in the Appendix.

2. NON-INFERIORITY HYPOTHESIS AND STATISTICAL INFERENCE

In this section, a fraction retention non-inferiority ratio hypothesis and the associated statistical test are established for a time-to-event endpoint. Analogous methods may be developed for other endpoints, such as odds ratio or relative risk.

A. Fraction Retention Hypothesis

Let the capital letters T, C, and P denote, respectively, the effects of experimental treatment, active control, and a reference "placebo" or "standard therapy" and HR for hazard ratio. The definition of a fraction retention of active control effect is given below.

$$\delta = \frac{[HR(P/C) - 1] - [HR(T/C) - 1]}{HR(P/C) - 1}. \tag{6}$$

The non-inferiority ratio hypothesis based on fraction retention can be formulated as follows.

$$H_0 : \delta \leq \delta_0 \quad \text{vs.} \quad H_a : \delta > \delta_0, \tag{7}$$

where $\delta_0 > 0$ is a specified fixed level of the relative fraction of retention desired.

The statistical inference for the fraction retention hypothesis in (7) is usually difficult because of a lack of acknowledge about the distributional properties of the associated test statistic. We propose a ratio test statistic in this paper to directly test the fraction retention ratio hypothesis in (7) with the desired power.

If $\delta_0 = 1$, then the fraction retention non-inferiority hypothesis in (7) becomes

$$H_0 : \delta \leq 1 \quad \text{vs.} \quad H_a : \delta > 1. \tag{8}$$

When we assumed that $HR(P/C) > 1$, i.e., the control treatment is truly effective, the non-inferiority ratio hypothesis in (8) is equivalent to

$$H_0 : HR(T/C) \geq 1 \quad \text{vs.} \quad H_a : HR(T/C) < 1. \tag{9}$$

The above hypothesis in (9) is a superiority hypothesis to compare the experimental treatment T to the control C.

If $\delta_0 = 0$, from the definition of fraction retention and the assumption that $HR(P/C) > 1$, the non-inferiority ratio hypothesis in (7) is equivalent to

$$H_0 : \frac{HR(T/C)}{HR(P/C)} \geq 1 \quad \text{vs.} \quad H_a : \frac{HR(T/C)}{HR(P/C)} < 1. \tag{10}$$

Under the constancy assumption that the control effect has not changed over time, the above hypothesis in (10) can be viewed as a surrogate superiority hypothesis to compare the experimental treatment T relative to a virtual placebo P:

$$H_0 : HR(T/P) \geq 1 \quad \text{vs.} \quad H_a : HR(T/P) < 1. \tag{11}$$

So, under the assumption that the control is truly effective and the assumption of constancy, the non-inferiority ratio hypothesis in (7) becomes a superiority hypothesis for testing the new treatment against the control when $\delta_0 = 1$. It becomes a superiority hypothesis for testing the new treatment against a virtual placebo when $\delta_0 = 0$. For appropriate choices of $0 < \delta_0 < 1$, the hypothesis in (7) becomes either an equivalence hypothesis for showing the equivalence between the new treatment and the control, or a non-inferiority hypothesis for showing the non-inferiority of the new treatment to the control.

B. Ratio Test Statistic

Let $\widehat{HR}(T/C)$ and $\widehat{HR}(P/C)$ be the observed hazard ratios of treatment effect relative to control and placebo effect relative to control, respectively. Obviously, the fraction retention δ can be estimated by

$$\hat{\delta} = \frac{\widehat{HR}(P/C) - \widehat{HR}(T/C)}{\widehat{HR}(P/C) - 1}. \tag{12}$$

Thus, a natural test statistic for the fraction retention hypothesis in (7) can be defined as follows.

$$Z^* = \frac{\hat{\delta} - \delta_0}{SE(\hat{\delta})}, \tag{13}$$

where $SE(\hat{\delta}) = \sqrt{\frac{1}{(\widehat{HR}(P/C)-1)^2}[s_x^2 + (1-\delta_0)^2 s_y^2]}$ is the standard error of $\hat{\delta}$ under the null hypothesis H_0 in (7), and where $s_x = SE(\widehat{HR}(T/C))$ and $s_y = SE(\widehat{HR}(P/C))$ are the standard errors of $\widehat{HR}(T/C)$ and $\widehat{HR}(P/C)$, respectively.

Since $\delta = 1 - \frac{HR(T/C)-1}{HR(P/C)-1}$ is a kind of ratio of two parameters $HR(T/C)$ and $HR(P/C)$ and Z^* is defined to directly test the ratio hypothesis in (7), Z^* is called a ratio test statistic. Hasselblad and Kong (2001) and Fisher et al. (2001) discussed similar ratio test statistics. However, they did not discuss the distributional properties of their ratio test statistics. In the later section of this paper, we will prove that Z^* has the asymptotic standard normal distribution under the null hypothesis H_0 in (7). To do so, we develop a tool based on the use of the logarithmic transformation and a sequence of interim statistics Z_k^* for the ratio hypothesis in (7), where k is a prespecified additive constant in each statistic. Then, Z^* is the limiting statistic based on Z_k^* such that Z^* has the optimal power. Since each Z_k^* is asymptotically normally distributed, Z^* has an asymptotic normal distribution under the null hypothesis H_0 in (7). Our simulation results show that Z^* converges to the standard normal with a relatively good convergence rate.

Since Z^* has the asymptotic standard normal distribution under the null hypothesis H_0 in (7), H_0 can be rejected for one-sided test if $z_0^* > c_{1-\alpha}$, where z_0^* is the observed value of Z^*, $c_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the critical value of normal test under the null hypothesis H_0 with the significance level α, and $\Phi(\cdot)$ is the cumulative probability function of the standard normal.

C. Asymptotic Normality of Z^*

In this section, we will derive the theoretical probability distribution of Z^*.

Theorem 1. *Z^* has the asymptotical standard normal distribution.*

$$Z^* = \frac{\hat{\delta} - \delta_0}{SE(\hat{\delta})} \longrightarrow N(0, 1), \tag{14}$$

where $SE(\hat{\delta}) = \sqrt{\frac{1}{(\widehat{HR}(P/C)-1)^2}[s_x^2 + (1-\delta_0)^2 s_y^2]}$.

The detailed proof of Theorem 1 is given in the Appendix. Though we have proved that Z^* has the asymptotic standard normal distribution under the null hypothesis H_0 in (7), for a small sample size, the convergence may be somewhat slow. In this section, we will report some simulation results for the normality of Z^*.

Table 1 summarizes the simulation results for the Xeloda trials (the detailed data can be found in the next section), where the simulation runs is 100,000 and p is the proportion of the simulation runs that passed the Shapiro-Wilk normality test. In the simulation study, two independent log-normal random variables for $\hat{\delta}$ are generated based on the values of the Xeloda trials and the assumptions of Theorem 1 that $\widehat{HR}(T/C)$ and $\widehat{HR}(P/C)$ are independent and have log-normal

Table 1 Simulation results for normality of Z^* (simulation runs = 100,000)

NOE	400	600	800	1000	1200	1400	1600	1800	2000
p	50.5%	68.2%	80.9%	88.9%	93.8%	96.6%	98.2%	99.1%	99.6%

distribution. From Table 1, if the number of events (NOE) is equal to or greater than 1000, we will have 89% probability of passing the Shapiro-Wilk normality test. Figure 1 shows the QQ-plot of Z^* for NOE = 1000. Table 1 and Figure 1 show that Z^* converges relatively quickly to the standard normal distribution.

Although Table 1 shows a large sample size (>1000 events), which would be required to ensure reasonable normality, in practice, non-inferiority trials often demand sample sizes that well exceed the numbers shown in this table. Therefore, the normality assumption should be readily satisfied.

D. Power and Sample Size

Let σ_0 and σ_a be the standard deviations of $\hat{\delta}$ under the null and alternative hypotheses in (7), respectively. At the design stage, the probability of type II error associated with Z^* test is given below.

$$\beta = \Phi\left(\frac{\delta_0 - \delta_a + \sigma_0 c_{1-\alpha}}{\sigma_a}\right). \tag{15}$$

To determine the sample size for a non-inferiority trial at the design stage, we assume that the standard deviations of $\hat{\delta}$ under the null and alternative hypotheses are the same, i.e., $\sigma_0 = \sigma_a$. Thus, from (15),

$$c_\beta = c_{1-\alpha} + \frac{\delta_0 - \delta_a}{\sigma_a},$$

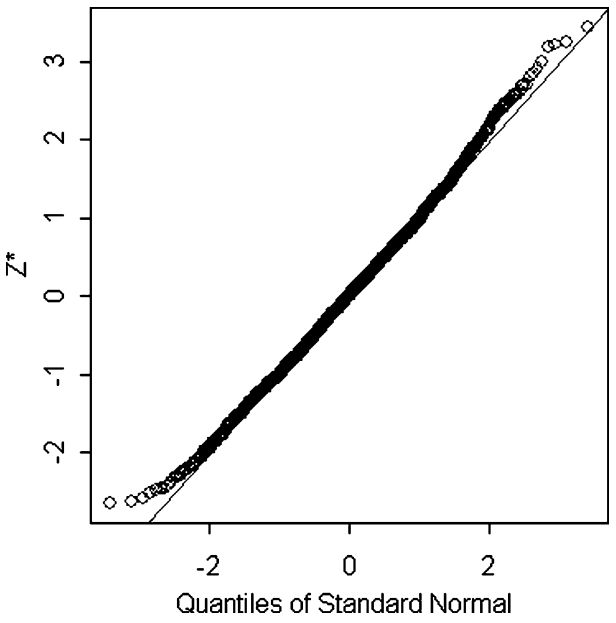

Figure 1 QQ Plot of Z^* with standard normal (NOE = 1000).

where $\sigma_a = \sqrt{\frac{1}{(HR(P/C)-1)^2}[\sigma_x^2 + (1-\delta_a)^2\sigma_y^2)]}$, and where σ_x^2 and σ_y^2 are the variances of $\widehat{HR}(T/C)$ and $\widehat{HR}(P/C)$, respectively. By the delta method, the variances can be approximately equal to

$$\sigma_x^2 \approx HR^2(T/C)\text{var}(\log \widehat{HR}(T/C)),$$

$$\sigma_y^2 \approx HR^2(P/C)\text{var}(\log \widehat{HR}(P/C)).$$

Thus, under the alternative hypothesis,

$$\sigma_x^2 + (1-\delta_a)^2\sigma_y^2 = \frac{(\delta_0 - \delta_a)^2(HR(P/C)-1)^2}{(c_\beta - c_{1-\alpha})^2}.$$

Let n_{x1} and n_{x2} denote the number of events in the experimental treatment and active control arms in the concurrent trial, respectively, and E denote the expected values. The asymptotic standard deviation of the log-hazard ratio in the concurrent trial can be estimated by $\sqrt{\frac{1}{E(n_{x1})} + \frac{1}{E(n_{x2})}}$. Fleming and Harrington (1991) gave a lower bound for the estimated standard error, which is $\sqrt{\frac{1}{E(n_{x1})} + \frac{1}{E(n_{x2})}} \geq \frac{2}{\sqrt{n_x}}$, where $n_x = n_{x1} + n_{x2}$. Thus, the sample size required for the current trial may be estimated by

$$n_x \geq \frac{4 \cdot (c_\beta - c_{1-\alpha})^2[(1-\delta_a)\widehat{HR}(P/C) + \delta_a]^2}{(\delta_a - \delta_0)^2(\widehat{HR}(P/C)-1)^2 - (c_\beta - c_{1-\alpha})^2(1-\delta_a)^2 s_y^2}. \tag{16}$$

An example about the sample size requirements can be found in the next section. It shows that the ratio approach based on Z^* requires a smaller sample size than the fixed margin approach based on the point estimate of the control effect.

3. EXAMPLE

In this section, we will use two Xeloda trials (FDA, 2001) as examples to illustrate the design and analysis of non-inferiority trial with the non-inferiority ratio hypothesis and statistical inference proposed in this paper. There were two randomized trials of about 600 patients, each comparing Xeloda with 5-FU + LV. For each trial, the efficacy criterion was a demonstration that Xeloda had a greater than 50% retention of the survival effect of 5-FU + LV relative to 5-FU alone.

Table 2 Survival results in two Xeloda clinical trials

Study	NOE	n	$HR(T/C)$	$SE(HR)$	95% CI(HR)	$SE(\log HR)$
SO14695	378	605	0.9964	0.0865	(0.8405, 1.1812)	0.0868
SO14796	386	602	0.9191	0.0797	(0.7754, 1.0893)	0.0867

Table 3 Meta-analysis results from 10 published papers

Study	$HR(P/C)$	$SE(HR)$	95% CI(HR)	$\log HR$	$SE(\log HR)$
10-papers	1.2638	0.0948	(1.0910, 1.4639)	0.2341	0.0750

Table 2 summarizes survival results for each trial based on the intent-to-treat populations.

To assess the control effect for Xeloda trials, 10 published papers based on randomized trials were selected, in which the control, 5-FU + LV, was compared with 5-FU alone. Although 5-FU alone may have a slightly better survival effect than a true placebo, there is no study reporting this conjecture. Table 3 summarizes survival results based on the meta-analysis of the 10 published papers.

For consistency, we use the following notations in the tables. $T =$ Xeloda, $C = 5\text{-}FU + LV$, and $P = 5\text{-}FU$. We also use the delta method to calculate the standard error of log-hazard ratio that $SE(\log HR) \approx HR^{-1}SE(HR)$.

In the FDA review report (2001), the efficacy criterion was a demonstration that Xeloda had a greater than 50% retention of the survival effect of 5-FU+LV relative to 5-FU alone, i.e., the null fraction retention hypothesis was $\delta \leq 50\%$ in each trial. The non-inferiority analyses are summarized in Table 4. The study power is calculated based on $\delta_a = 1$, which corresponds to equivalence between the experimental treatment and active control in the alternative hypothesis. The results of efficacy analysis are very close to the FDA review report. However, the study power of the proposed ratio test is much larger than Rothmann's method.

Table 5 lists the required number of events for different alternative hypotheses where the control effect ($HR(P/C)$) is assumed to be the estimated value 1.2638 and the associated standard error is assumed to be the estimated value 0.0948. Table 5 also summarizes the sample size requirement by different approaches: fixed margin approach based on the estimates of lower 95% confidence limit (95% LCL), fixed margin approach based on the point estimate of the control effect, Rothmann's method based on the test statistic Z_R^*, and the ratio test based on the test statistic Z^*. It shows that both fixed margin approaches based on the estimates of 95% LCL and Rothmann's method based on Z_R^* require a huge sample size. The ratio approach based on Z^* requires a much smaller sample size than the fixed margin approach

Table 4 Non-inferiority analysis for two Xeloda clinical trials ($\delta_0 = 50\%$, $\delta_a = 1$)

Study	z_0^*	$\hat{\delta}$	p-value	Power	95% CI of $\hat{\delta}$
SO14695	1.3741	101.4%	0.0847	26.17%	(39.9%, 163%)
SO14796	2.2957	130.7%	0.0109	62.34%	(72.9%, 188%)

Table 5 Number of events required for Xeloda trials (power = 80%, $\alpha = 0.05$)

		$\delta_0 = 50\%$			
δ_a	$HR(T/C)$	Fixed margin (95% LCL)	Fixed margin (point est.)	Rothmann method	Ratio test
0.90	1.024	24984	3278	15299	3171
0.95	1.012	19652	2558	7641	2316
1	1	15846	2045	4799	1805
1.05	0.988	13037	1668	3374	1465
1.10	0.977	10905	1384	2537	1222
1.15	0.965	9249	1163	1995	1041
1.20	0.954	7939	990	1619	901
1.25	0.943	6884	851	1346	789
1.30	0.932	6022	737	1140	697
1.35	0.921	5310	644	980	621
1.40	0.894	4715	567	853	557

based on the point estimate of the control effect. More discussion about these approaches can be found in the next section.

4. DISCUSSION

In the design of a non-inferiority trial, either a fixed margin or a ratio hypothesis may be formulated. A fixed margin hypothesis may be preferred by some because it can be easily understood and a standard statistical test can be applied. However, the question is how to select an appropriate margin. If the margin is selected inappropriately, one may conclude that a treatment that is actually worse than placebo is non-inferior to the active control. Now to choose the fixed margin appropriately, one needs to have some information on the control effect. If the margin is defined in terms of the estimate of the control effect based on some historical studies, then the standard statistical inference may not be valid, since the margin itself is variable. This is precisely the reason that led to the development of Rothmann's method. However, as discussed earlier, Rothmann's method has its own share of issues. The first issue is the difficulty of interpreting the fraction retention because it is defined in terms of log hazard. The second issue is its lack of power, resulting in impractical demands on sample size. Our proposed method defines the fraction retention directly in terms of the hazard, and thus it becomes easier to interpret the level of retention desired. In most commonly encountered situations, our proposed method enjoys a significant reduction in the sample size required when compared to either the fixed margin approach or Rothmann's method. Even then, the sample size required is still very large.

As we have discussed earlier, the ratio hypothesis (7) becomes a superiority hypothesis for testing the new treatment against the control when $\delta_0 = 1$. It becomes a superiority hypothesis for testing the new treatment against a virtual placebo when $\delta_0 = 0$. For appropriate choices of $0 < \delta_0 < 1$, the hypothesis in (7) becomes either an equivalence hypothesis for showing the equivalence between the new treatment and the control, or a non-inferiority hypothesis for showing the non-inferiority

of the new treatment to the control. Based on the Xeloda examples, it seems to suggest that $\delta_0 = 0.5$, i.e., a 50% retention, would correspond to a requirement for demonstrating equivalence, while $\delta_0 = 0.35 - 0.40$ would correspond to a requirement for demonstrating non-inferiority. In view of the power calculation, the 50% retention may be a reasonable level for demonstrating equivalence, while a 35% to 40% retention would be appropriate for demonstrating non-inferiority. With these levels, it would become possible to carry out non-inferiority trials due to the reasonable sample sizes required. A subsequent paper should be able to provide some insight and guidance in this regard.

There is one other important point that needs to be mentioned. As we indicated earlier, in the ratio hypothesis (7), $\delta_0 = 0$ would correspond to a surrogate superiority hypothesis (10) for demonstrating the superiority of the new treatment to a virtual placebo. This surrogate superiority hypothesis is the active control trial equivalent of the traditional placebo control trial where the new treatment is asked to show superiority to the placebo. This kind of surrogate hypothesis would permit the demonstration of the effectiveness of a new treatment in an active control trial. Of course, it presumes that there are adequate and well-controlled historical studies available demonstrating the effectiveness of the control. The necessary statistical tools for testing the surrogate superiority hypothesis (against a virtual placebo) will be discussed in a separate paper.

Rothmann has discussed the relationship between the fraction retention method and the fixed non-inferiority margin. He demonstrated that for a given level of retention and historical data on the control, there is a corresponding margin that can be calculated that would give rise to a corresponding fixed margin hypothesis. This approximate fixed margin can similarly be derived from our ratio test as follows.

From the definitions of Z^* and $\hat{\delta}$, we have

$$Z^* = \frac{-(\widehat{HR}(T/C) - 1) + (1 - \delta_0)(\widehat{HR}(P/C) - 1)}{2\sqrt{s_x^2 + (1 - \delta_0)^2 s_y^2}}, \tag{17}$$

where s_x is the standard error of $(\widehat{HR}(T/C) - 1)$.

Thus, for a test with one-sided significance level α and a 1:1 randomization, by the convergence theory, the corresponding two confidence interval testing procedure that rejects the fixed margin null hypothesis when the $100(1 - 2\alpha)\%$ two-sided confidence interval for $HR(T/C)$ lies entirely beneath the cutoff, which is approximating given by the following

$$1 + c_{1-\alpha} s_x + (1 - \delta_0)(\widehat{HR}(P/C) - 1) + 2c_\alpha \sqrt{s_x^2 + (1 - \delta_0)^2 s_y^2}. \tag{18}$$

As discussed above, the fixed margin non-inferiority hypothesis using this margin derived from the ratio test can be viewed roughly as the fixed margin equivalent of our ratio test. However, as we have mentioned earlier, this margin is not really a fixed constant since it depends on the trial data. It should also be noted that this rough equivalence can only be done in the setting of time-to-event endpoint due to the availability of the approximation used for s_x.

APPENDIX. PROOF OF THEOREM 1

Before we derive the probability distribution of Z^*, we need to define an interim statistic based on the logarithmic transformation.

$$Z_k^* = \frac{\log(\hat{\delta}+k)^2 - \log(\delta_0+k)^2}{SE(\log(\hat{\delta}+k)^2)}, \tag{19}$$

where $SE(\log(\hat{\delta}+k)^2) = \sqrt{\frac{4}{(\delta_0+k)^2(\widehat{HR}(P/C)-1)^2}[s_x^2 + (1-\delta_0)^2 s_y^2]}$.

In the statistic Z_k^*, k is a specified additive constant that is usually chosen as a nonnegative integer. Berry (1987) proposed the transformation $\log(x+k)$ for a parametric analysis such that the subsequent analysis is robust against heavy-tailed alternatives. In this paper, we use Z_k^* to derive the probability distribution of the ratio test statistic Z^*. For this purpose, Z_k^* is only an interim statistic. However, Z_k^* may also be used for the testing of the ratio hypothesis in (7) where the power of Z_k^* is monotonically increasing and is less than the power of Z^*. In order to focus on Z^*, the discussion of Z_k^* is not being considered in this paper.

Lemma 1. *For a fixed k, $\log(\hat{\delta}+k)^2$ has an asymptotic normal distribution and*

$$Z_k^* = \frac{\log(\hat{\delta}+k)^2 - \log(\delta_0+k)^2}{SE(\log(\hat{\delta}+k)^2)} \longrightarrow N(0,1), \quad (O(N^{-1/2}))$$

where $SE(\log(\hat{\delta}+k)^2) = \sqrt{\frac{4}{(\delta_0+k)^2(\widehat{HR}(P/C)-1)^2}[s_x^2 + (1-\delta_0)^2 s_y^2]}$ is the standard error of $\log(\hat{\delta}+k)^2$ under the null hypothesis H_0 in (7).

Hence, Z_k^* has the asymptotic standard normal distribution under the null hypothesis H_0 in (7). From the definitions of Z_k^* and Z^*, we have the following result.

Lemma 2. *Z^* is the limit of Z_k^* when $k \to \infty$.*

$$Z_k^* = \frac{1}{A}\log\left(\frac{\hat{\delta}+k}{\delta_0+k}\right)^{\delta_0+k} \longrightarrow \frac{\hat{\delta}-\delta_0}{A} = Z^*, \quad \text{when } k \to \infty$$

where the convergence is under the probability distribution and A is the standard error of $\hat{\delta}$

$$\begin{aligned}\frac{dg'}{dt}(\theta)\Sigma(\theta)\frac{dg}{dt}(\theta) = {} & \frac{4}{(\delta+k)^2(HR(P/C)-1)^2} \\ & \times[\sigma_1^2 HR^2(T/C) - 2\rho\sigma_1\sigma_2(1-\delta)HR(T/C)HR(P/C) \\ & + \sigma_2^2(1-\delta)^2 HR^2(P/C)],\end{aligned}$$

where ρ is the correlation coefficient between $\hat{\theta}_1$ and $\hat{\theta}_2$.

Since the control treatments are different in the concurrent trial and non-concurrent trials, $\hat{\theta}_1$ and $\hat{\theta}_2$ are usually assumed to be independent. That is, $\rho = 0$ in practice and then, by Slutsky's theorem and the definition of rate of convergence,

$$\frac{\log(\hat{\delta}+k)^2 - \log(\delta+k)^2}{\sqrt{\frac{dg}{dt}'(\theta)\Sigma(\theta)\frac{dg}{dt}(\theta)}} \longrightarrow N(0,1), \quad (O(N^{-1/2})).$$

Since $\frac{dg}{dt}'(\theta)\Sigma(\theta)\frac{dg}{dt}(\theta)$ is the variance of $\log(\hat{\delta}+k)^2$ and

$$\frac{\sqrt{\frac{dg}{dt}'(\theta)\Sigma(\theta)\frac{dg}{dt}(\theta)}}{\sqrt{\frac{dg}{dt}'(\hat{\theta})\Sigma(\hat{\theta})\frac{dg}{dt}(\hat{\theta})}} \longrightarrow 1 \qquad \text{with probability 1,}$$

the result in Lemma 1 follows from Slutsky's theorem under the null hypothesis H_0 in (7).

REFERENCES

Berry, D. (1987). Logarithmic transformations in ANOVA. *Biometrics* 43:439–456.

Brittain, E., Lin, D. (2003). *Non-Inferiority Trials of Antibiotic Therapy*. FDA-Industry Workshop.

Campbell, G., Yue, L. (2003). *Active Control Non-Inferiority Studies in Medical Devices*. FDA-Industry Workshop.

Chi, G. Y. H., Chen, G., Rothmann, M., Li, N. (2003). Active control trials. *Encyclopedia of Biopharmaceutical Statistics* 9–15.

Chen, G., Wang, Y. C., Chi, G. Y. H. (2004). Hypotheses and type I error in active control non-inferiority trials. *Journal of Biopharmaceutical Statistics* 14(2):301–303.

Ellenberg, S. S., Temple, R. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments—Part 2: Practical issues and specific cases. *Annals of Internal Medicine* 133:464–470.

FDA. (2001). *Medical-Statistical Review for Xeloda (NDA 20-896)*. FDA Division of Freedom of Information, Rockville, MD, dated 23 April, 2001.

Fisher, L. D., Gent, M., Buller, H. R. (2001). How would a new agent compare with placebo? a method illustrated with clopidogrel, aspirin, and placebo. *American Heart Journal* 141:26–32.

Fleming, T. R., Harrington, D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Hasselblad, V., Kong, D. F. (2001). Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal* 35:435–449.

Hauck, W., Anderson, S. (1999). Some issues in the design and analysis of equivalence trials. *Drug Information Journal* 33:109–118.

Hauschke, D., Kieser, M., Diletti, E., Burke, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distribution data. *Statistics in Medicine* 18:93–105.

Holmgren, E. B. (1999). Establishing equivalence by showing that a prespecified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* 9(4):651–659.

Hung, H. M. J., Wang, S. J., Tsong, Y., Lawrence, J., O'Neill, R. T. (2003). Some fundamental issues for non-inferiority testing in active controlled trials. *Stat. Medicine* 22:213–225.

Koch, G. G., Tangen, C. M. (1999). Nonparametric analysis of covariance and its role in non-inferiority clinical trials. *Drug Information Journal* 33:1145–1159.

Ng, T. H. (1993). A specification of treatment difference in the design of clinical trials with active controls. *Drug Information Journal* 27:705–719.

Pigeot, I., Schafer, J., Rvhmel, J., Hauschke, D. (2001). Assessing the therapeutic equivalence of two treatments in comparison with a placebo group. *Statistics in Medicine* 22:883–899.

Rothmann, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R., Tsou, H. H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 22:239–264.

Temple, R. (1996). Problem in interpreting active control equivalence trials. *Accountability in Research* 4:267–275.

Temple, R., Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments—Part 1: Ethical and scientific issues. *Annals of Internal Medicine* 133:455–463.

Wang, S. J., Hung, H. M. (2003). TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine* 22:227–238.

Walton, M., Gupta, S. (2003). *Non-Inferiority Margin Setting: Thrombolytics for Acute MI Example*. FDA-Industry Workshop.

White, H. D. (1998). Thrombolytic therapy and equivalence trials. *J. Am. Coll. Cardiol.* 31:494–496.

World Medical Association. (2000). Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Journal of the American Medical Association* 284:3043–3045.