Open Journal of Statistics, **2013, 3, 55-64**
http://dx.doi.org/10.4236/ojs.2013.32008 Published Online April 2013 (http://www.scirp.org/journal/ojs)

New Tests for Assessing Non-Inferiority and Equivalence from Survival Data

Kallappa M. Koti
United States Food and Drug Administration, Silver Spring, USA
Email: Kallappa.Koti@fda.hhs.gov

Received October 5, 2012; revised November 10, 2012; accepted November 20, 2012

ABSTRACT

We propose a new nonparametric method for assessing non-inferiority of an experimental therapy compared to a standard of care. The ratio μ_E/μ_R of true median survival times is the parameter of interest. This is of considerable interest in clinical trials of generic drugs. We think of the ratio m_E/m_R of the sample medians as a point estimate of the ratio μ_E/μ_R. We use the Fieller-Hinkley distribution of the ratio of two normally distributed random variables to derive an unbiased level-α test of inferiority null hypothesis, which is stated in terms of the ratio μ_E/μ_R and a pre-specified fixed non-inferiority margin δ. We also explain how to assess equivalence and non-inferiority using bootstrap equivalent confidence intervals on the ratio μ_E/μ_R. The proposed new test does not require the censoring distributions for the two arms to be equal and it does not require the hazard rates to be proportional. If the proportional hazards assumption holds good, the proposed new test is more attractive. We also discuss sample size determination. We claim that our test procedure is simple and attains adequate power for moderate sample sizes. We extend the proposed test procedure to stratified analysis. We propose a "two one-sided tests" approach for assessing equivalence.

Keywords: Right-Censored Data; Kaplan-Meier Estimate; Bootstrap Standard Error; Generic Drugs

1. Introduction

Non-inferiority and equivalence trials aim to show that the experimental therapy is not clinically worse than (non-inferiority) or clinically similar to (equivalence) an active control therapy. As the statistical formulation is one-sided, non-inferiority trials are also called one-sided equivalence trials. ICH E10 [1] is an authentic and official guidance document on the choice controls in non-inferiority clinical trials. The active control, which is also called a reference, is usually a standard of care. As noted in [1], most active-control equivalence trials are really non-inferiority trials intended to establish the efficacy of a new therapy. A non-inferiority trial is conducted to evaluate the efficacy of an experimental therapy compared to an active control when it is hypothesized that the experimental therapy may not be superior to a proven effective therapy, but is clinically and statistically not inferior in effectiveness. If the experimental therapy has a better safety profile, and/or easier to administer, and/or costs less, then non-inferiority trials are considered appropriate [2].

Confidence intervals on hazard ratios are used to assess equivalence and non-inferiority from survival data. The concept of hazard ratio is elusive. Clinicians find it hard to understand. Koch [3] says that though it is straightforward to construct confidence intervals on hazard ratios, it can be awkward to interpret. Wellek [4] proposed a log-rank test for equivalence of two survivor functions. According to Wellek, the survivor functions are considered equivalent if the absolute difference between the two survival curves is less than a pre-specified margin $\delta(>0)$ over the whole range of values of event-time. His test is carried out in terms of the regression coefficient for a dummy covariate indexing the trial arms. Though Wellek's paper is remarkable in its technical content, the test procedure is not used in practice. A possible reason is that his definition of equivalence criterion is conceptually difficult for clinicians to understand. Moreover, this formulation of the problem requires that the survival curves belong to the same proportional hazards model. The proportional hazards assumption is often inappropriate. We would like to point out that if the proportional hazards assumption holds good, the tests for

non-inferiority (and equivalence) in terms of medians would be more attractive.

Because the distribution of survival times tends to be positively skewed, the median is the preferred summary measure of the location of the distribution. Also, the median is straightforwardly informative to the clinicians. Efron [5] said it very nicely—"The median is often favored as a location estimate in censored data problems because, in addition to its usual advantage of easy interpretability, it least depends upon the right tail of the Kaplan-Meier curve, which can be highly unstable if censoring is heavy." Simon [6] emphasizes the importance of confidence intervals on median survival times. He writes: "For exponential survival distributions, the hazard ratio equals the ratio of medians. Exponential survival means that the survival curve is a straight line on a semi-logarithmic scale (log survival probability over time). Because exponential distributions are good approximations to the survival curves seen in many kinds of advanced cancer, confidence intervals for the hazard ratio are often interpreted as confidence intervals for the ratio of medians." Simon also explains how to calculate a confidence interval on the ratio of median survivals when the survival distributions are exponential. As a result, it has become a common practice in clinical trial study reporting to give point and interval estimates for the median survival time. This motivated us to consider testing for equivalence and non-inferiority of an experimental therapy compared to a reference therapy in terms of their median survival times. As assessing non-inferiority in terms of the difference between median survival times is trivial, we focus on their ratio.

Rubinstein *et al*. [7] were probably the first to consider the problem of testing the null hypothesis that the median survival times are equal against an alternative that the median survival time for the experimental treatment exceeds that of the control arm. They assumed exponential distributions for survival data. Britsol [8] presents a modification to Rubinstein's procedure for situations where it is desired to show that the experimental treatment is not much worse than the control. As noted by Berger and Hsu [9], and Hauschke and Hothorn [10], testing for non-inferiority in terms of the ratio of the averages often reflects clinical rationale rather than the difference between the averages. Bristol wants to test the null hypothesis that the ratio of medians is less than or equal to a fixed margin δ against the alternative that the ratio exceeds δ. To simplify the matter, he assumes that failure times have exponential distributions. Bristol's real interest is in testing the ratio hypothesis H_0^1 stated in (3.1) below in Section 3. However, he uses log transformation of the ratio to derive an asymptotic test. We circumvent this problem by introducing the Fieller-Hinkley (hereafter abbreviated as F-H) distribution on the ratio of two normally distributed random variables. Moreover, we don't assume failure times to follow exponential or some other parametric distributions.

2. One Sample Survival Model, Median Estimate and Standard Error

We develop the tests under the frame work of a randomly right-censored survival model. We assume that $Y_1, Y_2, \cdots, Y_n$ are iid random variables with a continuous distribution function F, and that F has a density f and median μ. These variables represent the event-times of the subjects under observation. Associated with each Y_i is an independent censoring variable C_i, which are assumed to be iid from a censoring distribution H. The data consist of n pairs (T_i, d_i), where T_i is either an observed failure-time Y_i or an observed censoring time C_i, and $d_i = I(Y_i = C_i)$. The basic quantity employed to describe time-to-event phenomenon is the survivor function $S(t) = 1 - F(t)$. The median survival time estimate is given by $m = \inf\{t : \hat{S}(t) \le 0.5\}$, where $\hat{S}(t)$ is the product-limit estimate of $S(t)$. That is, the median survival time is estimated from the product-limit estimate to be the first time that the survival curve falls to 0.5 or below. The sample median m is asymptotically normally distributed with mean μ. The variance $\sigma^2(m)$ of m is mathematically intractable. The SAS lifetest procedure provides an estimate of survivor function accompanied by survival standard error [11]. By default, the SAS lifetest procedure uses the Kaplan-Meier method. It also produces a point estimate of the median μ of F and the 95% confidence interval-derived by Brookmeyer and Crowley [12]. Brookmeyer and Crowley obtained the confidence intervals by inverting a generalization of the sign test for censored data. They did not need the standard error of the sample median. Obviously, the SAS lifetest procedure does not provide the standard error of the sample median m. One form of the asymptotic variance of median m is

$$\sigma^2(m) = \left[\hat{f}(\mu)\right]^{-2} \times \operatorname{var}\left[\hat{S}(\mu)\right], \qquad (2.1)$$

where $\hat{S}(m)$ is found using the Greenwood's formula [13]. A slightly different version of $\sigma^2(m)$ is provided in [14]:

$$\sigma^2(m) = n^{-1}\left[f(\mu)\right]^{-2} \cdot \left[\left\{1 - F(\mu)\right\}^2 \int_0^{\mu} \frac{\mathrm{d}F}{(1-F)(1-H)}\right] \qquad (2.2)$$

As f is unknown, the variance $\sigma^2(m)$ given either in (2.1) or (2.2) becomes useless in estimating the population median time μ [15]. We propose to estimate the standard error of m using the Efron's bootstrap [5], which

does not make any distributional assumptions. In a single sample setting, Efron's bootstrap may be described as follows. We draw a bootstrap sample $\left(Y_1^*, C_1^*\right), \left(Y_2^*, C_2^*\right), \cdots, \left(Y_n^*, C_n^*\right)$ by independent sampling n times with replacement from F and calculate the median $m^* = m\left(\text{data}^*\right)$. We repeat this independently B times, obtaining B medians: $m^{*1}, m^{*2}, \cdots, m^{*B}$. An estimated variance of the sample median time m is

$$\hat{\sigma}^2_{\text{BOOT}} = \frac{1}{B-1}\left[\sum_{j=1}^{B}\left(m^{*j}\right)^2 - \left(\sum_{j=1}^{B} m^{*j}\right)^2 \Big/ B\right] \quad (2.3)$$

One may set B equal to 1000. This is called "model-free" or the Efron's bootstrap procedure II. The University of Texas at Austin [16] has provided some introductory SAS codes needed to resample a SAS dataset.

Efron [5] states: the bootstrap estimate $\hat{\sigma}_{\text{BOOT}}$ given in (2.3) is a consistent estimate, but σ in (2.1) or in (2.2) itself may be meaningless. Therefore, we assume that $\hat{\sigma}^2_{\text{BOOT}}$, which does not depend on either f or μ is a viable substitute for $\sigma^2(m)$. Thus, we work under the notion that the sample median time m is asymptotically normally distributed with mean μ and variance $\hat{\sigma}^2_{\text{BOOT}}$. We suppress the subscript BOOT of the estimated variance in (2.3). In fact, Keaney and Wei [17], among others, have used bootstrap to find the standard error of m.

What is an indication of an unstable median or heavy censoring is a crucial question. As observed in [12], if the survival curve is relatively flat in the neighborhood of 50% survival, there can be great deal of variability in the estimated median. It would be more appropriate to cite a confidence interval for the median. We propose a simple rule of thumb. If the upper limit of a 95% confidence interval on median is not available, one may conclude that median is unstable and/or censoring is heavy. Therefore, the proposed tests should work efficiently when the Brookmeyer-Crowley upper limit of a 95% confidence interval on median is available. This also minimizes the number of bootstrap samples whose Kaplan-Meier curves do not reach 0.5 survival probability. In addition, asymptotic normality requires that $m > 2\hat{\sigma}$.

3. Null and Alternative Hypotheses

Let T_E and T_R denote the times to event for the experimental and reference treatment groups, respectively. We use S_E and S_R to denote the survival functions, and μ_E and μ_R to denote the medians of T_E and T_R, respectively. Depending on the application one may test

$$H_0^1 : \mu_E/\mu_R \le \delta_L \text{ vs. } H_A^1 : \mu_E/\mu_R > \delta_L \quad (3.1)$$

Here $\delta_L < 1$ and large median values point to large positive effects. For example, the null and alternative hypotheses in (3.1) are appropriate if non-inferiority as measured by the overall survival of patients is desired. In some other applications, small median values may point to large positive effects, in which case, for proving non-inferiority, one may test

$$H_0^2 : \mu_E/\mu_R \ge \delta_U \text{ vs. } H_A^2 : \mu_E/\mu_R < \delta_U \quad (3.2)$$

where $\delta_U > 1$. For example, if duration of anemia (or time to response) is the clinical endpoint, it is appropriate to consider the null and alternative hypotheses in (3.2). Here H_A^1 and H_A^2 indicate that the experimental therapy is not inferior to the reference therapy. The lower and upper bounds δ_L and δ_U defining non-inferiority are called non-inferiority margins. The selection of non-inferiority margin δ_L (or δ_U) depends upon a combination of statistical reasoning and clinical judgment. For a discussion on the choice of a non-inferiority margin, reference is made to ICH-E10 document [1]. For example, testing

$$H_0^1 : \mu_E/\mu_R \ge 0.8 \text{ or } H_0^2 : \mu_E/\mu_R \le 1.25 \quad (3.3)$$

is of considerable interest in clinical trials of generic drugs. Henceforth, we assume that two independent sample $\left(T_{E,1}, T_{E,2}, \cdots, T_{E,n_E}\right), \left(T_{R,1}, T_{R,2}, \cdots, T_{R,n_R}\right)$ of possibly right-censored event-times are given. We use $\boldsymbol{T}$ to represent the data. The sample size n_E and n_R are sufficiently large. The censoring proportion, in each arm, is moderate. That is, the trial is designed to have long enough follow-up time so that more than one half of the subjects in both arms had the event. Let $\hat{S}_E$ and $\hat{S}_R$ denote the product-limit survival estimates and m_E and m_R denote the median time estimates for the experimental and reference groups, respectively. The sample medians m_E and m_R are independently asymptotically normally distributed with means μ_E and μ_R, and variances σ_1^2 and σ_2^2, respectively. As mentioned in Section 2, we assume that the bootstrap variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ given by (2.3) are the de facto variances of m_E and m_R, respectively. The proportional hazards assumption is not required. However, we assume that the each treatment group has survival curve that is not relatively flat in the neighborhood of 50 percent survival. We also assume that each median estimate is at least two times larger than its standard error. Then the ratio $W = m_E/m_R$ follows the F-H distribution that is briefly described in the next section.

4. Fieller-Hinkley Distribution

Let X_1 and X_2 be normally distributed random variables with means θ_i, variances $\sigma_i^2\,(i = 1, 2)$ and correlation coefficient ρ. Let $W = X_1/X_2$. Fieller [18] obtains the probability density function $g(w)$ of W. Hinkley [19] derives the cumulative distribution function $G(w)$ of W. We have not shown $g(w)$ and $G(w)$

here due to lack of space. As a special case, Hinkley has shown that as $\theta_2/\sigma_2 \to \infty$, that is, as $P(X_2 > 0) \to 1$,

$$G(w) \to G^+(w) = \Phi\left[\frac{\theta_2 w - \theta_1}{\sigma_1\sigma_2 a(w)}\right], \quad a(w) = \left(\frac{w^2}{\sigma_1^2} - 2\frac{\rho w}{\sigma_1\sigma_2} + \frac{1}{\sigma_2^2}\right)^{1/2} \tag{4.1}$$

where Φ denotes the standard normal distribution function. In what follows, we consider the case where X_1 and X_2 are statistically independent, and therefore, we set $\rho = 0$. Note that the argument in Φ may be written as $\theta_2(w-\delta)/\sigma_1\sigma_2 a(w)$, where $\delta = \theta_1/\theta_2$. The probability density function corresponding to G^+, when $\rho = 0$, is

$$g^+(w) = \frac{\theta_2\left(\sigma_1^2 + \sigma_2^2\delta w\right)}{\left[\sigma_1\sigma_2 a(w)\right]^3} \times \phi\left(\frac{\theta_2(w-\delta)}{\sigma_1\sigma_2 a(w)}\right), w > 0,$$

where ϕ denotes the standard normal density function.

The distribution G^+ is unimodal but not necessarily symmetric. It has a median equal to θ_1/θ_2. The superscript + in G^+ refers to W being a positive valued random variable. As the ratio of median survival times is always positive, we suppress the superscript.

Koti used the F-H distribution to derive non-inferiority tests under analysis of variance setting [20]. Koti also used the F-H distribution to derive tests for null hypothesis of non-unity ratio of proportions [21]. In this paper, his test procedure is extended to survival data analysis. We think of the ratio $W = m_E/m_R$ as a point estimate of the ratio μ_E/μ_R and we intend to use the distribution G of the ratio W to make inference on μ_E/μ_R. As usual, w denotes an observed value of W. We regard the variances σ_1^2 and σ_2^2 as nuisance parameters. In what follows, we replace σ_1^2 and σ_2^2 by their bootstrap estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, respectively.

5. Test for the Lower Inequality

In this section we consider testing the null hypothesis H_0^1 against the alternative hypothesis H_A^1, which are stated in (3.1). Under the null hypothesis $H_0^1: \mu_E/\mu_R = \delta_L$, the distribution function of $W = m_E/m_R$, the ratio of sample medians, is given by

$$G\left(w \middle| H_0^1\right) = \Phi\left[\frac{\mu_R(w-\delta_L)}{\hat{\sigma}_1\hat{\sigma}_2 a(w)}\right], \quad a(w) = \left(\frac{w^2}{\hat{\sigma}_1^2} + \frac{1}{\hat{\sigma}_2^2}\right)^{1/2} \tag{5.1}$$

Intuitively, H_0^1 should be rejected in favor of H_A^1 for large observed values of W. We reject H_0^1 in favor of H_A^1 if $W > w'$, where

$$P\left(W > w' \middle| H_0^1\right) = 1 - \Phi\left[\frac{\mu_R(w'-\delta_L)}{\hat{\sigma}_1\hat{\sigma}_2 a(w')}\right] = \alpha.$$

We need to find a cutoff point w' that satisfies the equation

$$\mu_R(w'-\delta_L)/\hat{\sigma}_1\hat{\sigma}_2 a(w') = z_{1-\alpha}, \tag{5.2}$$

where z_a is the 100a-th percentile of the standard normal distribution. The cutoff point w' satisfying (5.2) defines the rejection region for a given value of μ_R. Note that δ_L is the median of W for all μ_R and the cutoff point $w' > \delta_L$ for $\alpha < 0.5$. To construct a test that has a significance level less than or equal to α for all μ_R, we proceed as follows. Calculate $100(1-\omega)$ percent confidence intervals on μ_E and μ_R, where $\omega \in [0.01, 0.05]$. Let (L_1, U_1) and (L_2, U_2) denote these confidence intervals on μ_E and μ_R, respectively. These confidence intervals should be as wide as possible. Let

$$\Theta_\omega = \left\{(\mu_E, \mu_R)' : L_1 < \mu_E < U_1, L_2 < \mu_R < U_2\right\} \tag{5.3}$$

We describe Θ_ω in (5.3) as a rectangular parameter space. Let $\ell_1 = \left\{(\mu_E, \mu_R)' : \mu_E = \delta_L \times \mu_R\right\} \subset \Theta_w$, and D_1^N denote the domain of the line ℓ_1. Here ℓ_1 represents the parameter space under the simple null hypothesis $H_0^1: \mu_E/\mu_R = \delta_L$. We assume that D_1^N is non-empty.

Consider $\mu_R^{(1)} < \mu_R^{(2)}$ where both $\mu_R^{(1)}$ and $\mu_R^{(2)}$ are in D_1^N and satisfy (5.2) for some $w^{(1)}$ and $w^{(2)}$. That is,

$$\frac{\mu_R^{(1)}\left(w^{(1)} - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right)} = \frac{\mu_R^{(2)}\left(w^{(2)} - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(2)}\right)} = z_{1-\alpha},$$

and $w^{(1)}, w^{(2)} > \delta_L$. It means that

$$\left(w^{(1)} - \delta_L\right)/\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right) > \left(w^{(2)} - \delta_L\right)/\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(2)}\right).$$

Now $(w-\delta_L)/\hat{\sigma}_1\hat{\sigma}_2\, a(w)$ increases as w increases. This implies that $w^{(1)} > w^{(2)}$ and

$$\mu_R^{(2)}\left(w^{(1)} - \delta_L\right)/\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right) > \mu_R^{(1)}\left(w^{(1)} - \delta_L\right)/\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right).$$

Therefore,

$$\Phi\left[\frac{\mu_R^{(2)}\left(w^{(1)} - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right)}\right] > \Phi\left[\frac{\mu_R^{(1)}\left(w^{(1)} - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right)}\right], \text{ and}$$

$$1 - \Phi\left[\frac{\mu_R^{(2)}\left(w^{(1)} - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right)}\right] < 1 - \Phi\left[\frac{\mu_R^{(1)}\left(w^{(1)} - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w^{(1)}\right)}\right] \tag{5.4}$$

This is graphically illustrated in **Figure 1**. Two F-H distribution functions with $\delta_L = 0.8$ are shown in **Figure 1**. The graph in solid line represents G with $\mu_R = 10.27$ and the other one represents G with $\mu_R = 12.0$. Here we have used $\hat{\sigma}_1 = 1.42$, and $\hat{\sigma}_2 = 2.787$. Note that in the upper half portion of **Figure 1**, the distribution function $G\left(w\middle|\mu_R = 10.27\right)$ runs below the distribution function $G\left(w\middle|\mu_R = 12.0\right)$. That is, for each x-coordinate $> \delta_L$, the y-coordinate for G with $\mu_R = 12.0$ is lower than the one for $\mu_R = 10.27$.

This is what is claimed in (5.4). The reader may note that $G\left(1.51\middle|\mu_R = 10.27\right) = 0.95$, and $G\left(1.51\middle|\mu_R = 12.0\right) < 0.95$. That is,

$1 - G\left(1.51\middle|\mu_R = c_1\right) = 0.05$, and

$1 - G\left(1.51\middle|\mu_R = c_2\right) < 0.05$.

Let $\tilde{\mu}_{R1}$ denote the smallest μ_R in D_1^N and $G_1(w) = G\left(w\middle|H_0^1, \tilde{\mu}_{R1}\right)$. Then from (5.4), it follows that $\tilde{w}_1 = G_1^{-1}(1-\alpha)$ defines the critical region. That is, reject H_0^1 if $W > \tilde{w}_1$. The significance level $\left[1 - G\left(\tilde{w}_1\middle|H_0^1\right)\right]$ is less than or equal to α for all $\mu_R \geq \tilde{\mu}_{R1}$. Therefore, the rule that rejects H_0^1 for $W > \tilde{w}_1$ is a level α test.

The cut off point $\tilde{w}_1$ can be determined as follows. Square both sides of Equation (5.2) with μ_R replaced by $\tilde{\mu}_{R1}$ and get a quadratic equation:

$a_1 w^2 + b_1 w + c_1 = 0$, where

$a_1 = z_{1-\alpha}^2 \hat{\sigma}_2^2 - \tilde{\mu}_{R1}^2, b_1 = 2 \times \delta_L \times \tilde{\mu}_{R1}^2$, and

$c_1 = z_{1-\alpha}^2 \hat{\sigma}_1^2 - \delta_L^2 \times \tilde{\mu}_{R1}^2$.

The roots of the quadratic equation are

$w_1\left(\tilde{\mu}_{R1}\right) = \left(-b_1 \pm \sqrt{b_1^2 - 4a_1c_1}\right) \Big/ 2a_1$. The root that is smaller than δ_L defines the critical region of the test. Alternatively, one may use the SAS PROBNORM for tabulating G_1 and find $\tilde{w}_1$.

5.1. *p*-Value and Power of the Test

The *p*-value for the test is

$$\alpha_L' = 1 - \Phi\left[\frac{\tilde{\mu}_{R1}\left(w_C - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w_C\right)}\right] \tag{5.5}$$

where $w_C = m_E / m_R$ is the observed ratio. The power of the proposed test is the probability that the null hypothesis H_0^1, will be rejected when the alternative hypothesis H_A^1, is true. We define the power function $\beta_L = \beta_L\left(\delta_A, \tilde{\mu}_{R1}\right)$ for a given alternative $\delta_A > \delta_L$, as

$$\beta_L = 1 - \Phi\left[\frac{\tilde{\mu}_{R1}\left(\tilde{w}_1 - \delta_A\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(\tilde{w}_1\right)}\right], \delta_A = \mu_E / \mu_R. \tag{5.6}$$

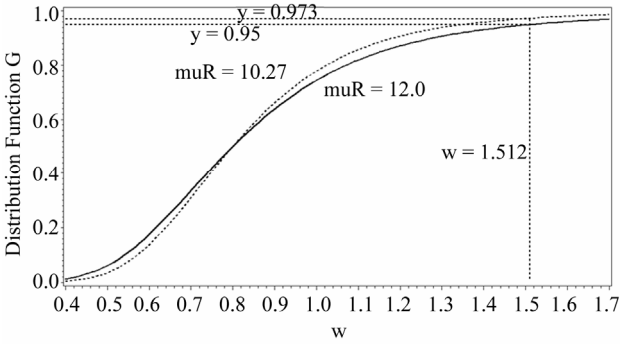

Figure 1. Two DFs of *W* both with a median of 0.8.

Usually, in designing a clinical trial, one aims to have a power over 0.5. Note that the power, for example, β_L in (5.6) exceeds 0.5 only if $\tilde{w}_1 < \delta_A$. For a given δ_a, it readily follows that $0.5 < \beta_L\left(\delta_a, \tilde{\mu}_{R1}\right) < \beta_L\left(\delta_a, \mu_R\right)$ for all $\tilde{\mu}_{R1} < \mu_R$. Therefore, the power β_L may be called the minimum power.

5.2. The Test Is Unbiased

Note that

$$\begin{aligned}\alpha &= 1 - \Phi\left[\frac{\tilde{\mu}_R\left(\tilde{w}_1 - \delta_L\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(\tilde{w}_1\right)}\right], \tilde{w}_1 > \delta_L \\ &< 1 - \Phi\left[\frac{\tilde{\mu}_{R1}\left(\tilde{w}_1 - \delta_A\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(\tilde{w}_1\right)}\right], \delta_A > \delta_L \\ &= \beta_L\left(\delta_A\right).\end{aligned}$$

That is, the type-I error probability is at most α and the power of the test is at least α. Thus, the test is unbiased.

6. Test for the Upper Inequality

Next, we discuss testing the null hypothesis H_0^2 against the alternative hypothesis H_A^2, which are stated in (3.2). The null hypothesis H_0^2 should be rejected in favor of H_A^2 for smaller observed values of the ratio $W = m_E / m_R$. As under $H_0^2 : \mu_E / \mu_R = \delta_U$, we set

$$G\left(w\middle|H_0^2\right) = \Phi\left[\frac{\mu_R\left(w - \delta_U\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w\right)}\right] = \alpha. \tag{6.1}$$

That is, we need to find a cutoff point w' that satisfies the equation

$$\tilde{\mu}_{R2}\left(w' - \delta_U\right) \Big/ \hat{\sigma}_1\hat{\sigma}_2 a\left(w'\right) = z_\alpha, \tag{6.2}$$

where $\tilde{\mu}_{R2}$ is the smallest μ_R in Θ_ω and $\mu_E = \delta_U \times \mu_R$. Let $\tilde{w}_2$ denote the solution of (6.2) that is less than δ_U. It follows that $G\left(\tilde{w}_2\middle|H_0^2\right) \leq \alpha$ for all $\mu_R \geq \tilde{\mu}_{R2}$.

p-Value and Power of the Test

The *p*-value for the test is

$$\alpha'_U = \Phi\left[\frac{\tilde{\mu}_{R2}\left(w_C - \delta_U\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w_C\right)}\right], \quad (6.3)$$

where $w_C = m_E/m_R$ is the observed ratio. The power function $\beta_U = \beta_U\left(\delta_A, \tilde{\mu}_{R2}\right)$ at $\delta_A = \mu_E/\mu_R, \delta_A < \delta_U$, is given by

$$\beta_U = \Phi\left[\frac{\tilde{\mu}_{R2}\left(\tilde{w}_2 - \delta_A\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(\tilde{w}_1\right)}\right], \delta_A = \mu_E/\mu_R \quad (6.4)$$

Note that the power, for example, β_U in (6.4) exceeds 0.5 only if $\tilde{w}_2 > \delta_A$. For a given δ_a, it readily follows that $0.5 < \beta_U\left(\delta_a, \tilde{\mu}_{R2}\right) < \beta_U\left(\delta_a, \mu_R\right)$ for all $\tilde{\mu}_{R2} < \mu_R$. Therefore, β_U in (6.4) may be called the minimum power. The test is unbiased.

7. Bootstrap Equivalent Confidence Intervals

In one sample case, for randomly right-censored survival model, Efron has considered using bootstrap to estimate the sampling distribution of $\sqrt{n}\left[\hat{S}(t) - S(t)\right]$, where n is the sample size [5]. He has demonstrated that the sampling distribution of $\sqrt{n}\left[\hat{S}(t) - S(t)\right]$ can be estimated by the distribution of $\sqrt{n}\left[\hat{S}^b(t) - \hat{S}(t)\right]$, where $\hat{S}^b$ denotes the bootstrap Kaplan-Meier estimate. See [5,22] for details on the method(s) of bootstrapping. Let $\hat{m}^b$ denote the Efron's bootstrap estimate of the median. Then Efron has shown that $\hat{m}^b - m$ has the same distribution under $\hat{F}$ as does $m - \mu$ under F [5]. But we know that m is asymptotically normally distributed with mean μ and variance σ_m^2. Therefore, it is reasonable to say that the bootstrap median estimate $\hat{m}^b$ is asymptotically normally distributed with mean equal to the sample median m and variance equal to $\hat{\sigma}_m^2$ [14]. We use this result to formulate a confidence interval based method for assessing non-inferiority of T_E compared to T_R.

Let $\hat{m}_E^b$ be the median estimate based on a bootstrap sample $\left(T_{E,1}^b, T_{E,2}^b, \cdots, T_{E,n_E}^b\right)$ taken with replacement from $\left(T_{E,1}, T_{E,2}, \cdots, T_{E,n_E}\right)$, and $\hat{m}_R^b$ denote the median estimate based on a bootstrap sample $\left(T_{R,1}^b, T_{R,2}^b, \cdots, T_{R,n_R}^b\right)$ taken with replacement from $\left(T_{R,1}, T_{R,2}, \cdots, T_{R,n_R}\right)$. By the above argument, it follows that $\hat{m}_E^b$ is asymptotically normally distributed with mean m_E and variance $\hat{\sigma}_1^2$, and $\hat{m}_R^b$ is asymptotically normally distributed with mean m_R and variance $\hat{\sigma}_2^2$. Note that $\hat{m}_E^b$ and $\hat{m}_R^b$ are independent. Therefore, the ratio $W^b = \hat{m}_E^b/\hat{m}_R^b$ has the distribution function

$$\hat{G}_B\left(w\right) = \Phi\left[\frac{m_R w - m_E}{\hat{\sigma}_1\hat{\sigma}_2 a\left(w\right)}\right], a\left(w\right) = \left(\frac{w^2}{\hat{\sigma}_1^2} + \frac{1}{\hat{\sigma}_2^2}\right)^{1/2}. \quad (7.1)$$

That is, we plug in the sample estimates of μ_E, μ_R, σ_1^2 and σ_2^2 in $G(w)$ of (4.1) to get an asymptotic distribution of the bootstrap ratio W^b. Note that the distribution function $\hat{G}_B\left(w\right)$ is completely specified.

Equivalence between the two treatments is often tested by the confidence interval approach, which consists of constructing a $100(1-2\alpha)$ percent confidence interval for the parameter of interest and comparing the constructed confidence interval with the pre-specified equivalence range [9]. In this paper, we use the distribution $\hat{G}_B\left(w\right)$ in (7.1) to obtain a $100(1-2\alpha)$ percent confidence interval for the ratio μ_E/μ_R for equivalence testing. A $100(1-2\alpha)$ percent confidence interval for the ratio μ_E/μ_R is given by

$$\mathrm{I}(\boldsymbol{T}) = \left(\hat{G}_B^{-1}(\alpha), \hat{G}_B^{-1}(1-\alpha)\right). \quad (7.2)$$

The interval in (7.2) may be obtained in two ways. One may tabulate $\left(w, \hat{G}_B\right)$ using SAS PROBNORM and locate the confidence limits. Alternatively, one may write down the quadratic equations of the type shown in (5.2) and (6.2) and solve them. See section 8 for illustration. If the constructed confidence interval $\mathrm{I}(\boldsymbol{T})$ falls within the equivalence limits $\left(\delta_L, \delta_U\right)$, then the two groups are considered equivalent. In order to demonstrate non-inferiority, this interval should lie entirely on the positive side of non-inferiority margin. That is, if the confidence interval in (7.2) excludes the non-inferiority margin, then non-inferiority is demonstrated.

8. Sample Size Determination

In the current setting, the standard error of sample median is not explicitly expressed in terms of the number of events. Therefore, we assume exponential model for sample size calculation. That is, we assume that T_E and T_R have exponential distribution with means λ_E^{-1} and λ_R^{-1}, respectively. Let $\hat{\lambda}_E$ and $\hat{\lambda}_R$ represent the maximum likelihood estimates of λ_E and λ_R, respectively The median time estimates are given by $m_E = \ln 2/\hat{\lambda}_E$ and $m_R = \ln 2/\hat{\lambda}_R$ [13]. Suppose that $r_E\left(< n_E\right)$ and $r_R\left(< n_R\right)$ are the numbers of observed event-times. For simplicity, we assume that $r_E = r_R = r$. The standard errors of m_E and m_R are given by $\hat{\sigma}_1 = \ln 2/\left(\hat{\lambda}_E\sqrt{r}\right)$ and $\hat{\sigma}_2 = \ln 2/\left(\hat{\lambda}_R\sqrt{r}\right)$, respectively. We describe the sample size determination for the test for the upper inequality. That is, we consider testing

$$H_0^2 : \mu_E/\mu_R \geq \delta_U \text{ vs. } H_A^2 : \mu_E/\mu_R < \delta_U.$$

8.1. Power Approach

We assume that m_R is given. That is, $\hat{\lambda}_R$ is known. To be consistent with H_0^2, we set $m_E = \delta_U m_R$. Therefore, we have $\hat{\sigma}_1 = \delta_U \hat{\sigma}_2$. The null distribution of W is given by

$$G\left(w \middle| H_0^2\right) = \Phi\left[\sqrt{r}\left(w-\delta_U\right) \Big/ \left(w^2+\delta_U^2\right)^{1/2}\right] \qquad (8.1)$$

We note that the distribution function G in (8.1) is a function of r and it does not explicitly depend on λ_R or μ_R. We find the cut-off point $\tilde{w}_2$ for a level α test either by solving $\sqrt{r}\left(w-\delta_U\right) = z_\alpha\left(w^2+\delta_U\right)^{1/2}$ or by tabulating G in (8.1). The power β at $\mu_E/\mu_R = 1$, as a function of r, is given by

$$\beta_r = \Phi\left[\sqrt{r}\left(\tilde{w}-1\right) \Big/ \left(\tilde{w}^2+1\right)^{1/2}\right] \qquad (8.2)$$

We calculate the optimal number of events r per arm, which yields a power of 0.8 for a test of size 0.05 by iteration. We start with $r = 30$. Find $\tilde{w}_2 = \tilde{w}_2(r)$, where $G\left(\tilde{w}_2 \middle| H_0^2\right) = \alpha$. Next, we calculate the power β_r given in (8.2). If the power is less than 0.8, we increase r, and repeat the procedure. We note that when the non-inferiority margin $\delta_U = 1.25$, the required number of events per arm is $r = 250$. Similarly, for $\delta_U = 1.5$, the number of events required per arm is $r = 77$. For testing $H_0^1 : \mu_E/\mu_R \le 0.8$ versus $H_A^1 : \mu_E/\mu_R > 0.8$, one needs $r = 251$ events per arm to achieve a power of 0.8 at $\mu_E/\mu_R = 1$.

8.2. Bootstrap Confidence Interval Approach

In this setting, the distribution function of $W = m_E/m_R$ is

$$\hat{G}_B(w) = \Phi\left[\sqrt{r}\left(\hat{\lambda}_E w - \hat{\lambda}_R\right) \Big/ \left(\hat{\lambda}_E^2 w^2 + \hat{\lambda}_R^2\right)^{1/2}\right].$$

To find an optimal sample size, we use $\hat{G}_B$ to find a $100(1-2\alpha)$ percent confidence interval. We set

$$\sqrt{r}\left(\hat{\lambda}_E w - \hat{\lambda}_R\right) \Big/ \left(\hat{\lambda}_E^2 w^2 + \hat{\lambda}_R^2\right)^{1/2} = \left|z_\alpha\right|$$

and solve for

$$w : \left(z_\alpha^2 \hat{\lambda}_E^2 - r\hat{\lambda}_E^2\right) w^2 + 2r\hat{\lambda}_R \hat{\lambda}_E w + z_\alpha^2 \hat{\lambda}_R^2 - r\hat{\lambda}_R^2 = 0.$$

The roots of this quadratic equation are given by $\tilde{w}(r) = \left(-B \pm \sqrt{B^2 - 4AC}\right) \Big/ 2A$, where

$A = z_\alpha^2 \hat{\lambda}_E^2 - r\hat{\lambda}_E^2, B = 2r\hat{\lambda}_R \hat{\lambda}_E$, and $C = z_\alpha^2 \hat{\lambda}_R^2 - r\hat{\lambda}_R^2$.

Let $\tilde{w}_1(r) = \left(-B + \sqrt{B^2 - 4AC}\right) \Big/ 2A$, and

$$\tilde{w}_2(r) = \left(-B - \sqrt{B^2 - 4AC}\right) \Big/ 2A.$$

Note that $\tilde{w}_1(r) < \tilde{w}_2(r)$ for $r > z_\alpha^2$. Then the interval $\left(\tilde{w}_1(r), \tilde{w}_2(r)\right)$ is a $100(1-2\alpha)$ confidence interval. The endpoints of the desired confidence interval are expressed in terms of r. Next, we propose to find r^*, an optimal r satisfying $\tilde{w}_2(r) - \tilde{w}_1(r) = d$ where d is a pre-specified constant. Ideally, the choice of d should depend on the width of $\left(\delta_L, \delta_L^{-1}\right)$ or $\left(\delta_U^{-1}, \delta_U\right)$. Note that the difference $\tilde{w}_2(r) - \tilde{w}_1(r) = d$ is written as $-\sqrt{B^2 - 4AC} = A \times d$. A closed form expression for r^* is not available. We note that $A < 0$ for $r > z_\alpha^2$. The optimal number of events per arm r^* is the smallest $r\left(> z_\alpha^2\right)$ such that $-\sqrt{B^2 - 4AC} - A \times d \ge 0$. The value of r^* is found by a simple computer search. We have provided values of r^* in **Tables 1** and **2** below when $\alpha = 0.025$ and $\alpha = 0.05$, respectively. In doing so, we have selected the pairs $\left(\lambda_E, \lambda_R\right)$ for which non-inferiority (or equivalence) investigation makes sense.

9. Stratified Analysis

In most phase 3 studies, stratified randomization is adopted. That is, subjects are grouped according to covariate values such as age group and baseline performance status prior to randomization and subjects are then randomized within strata.

Within each stratum, a separate randomization sequence to allocate subjects to treatment groups is used. In this section, we extend the above test procedure to clinical trials, which consist of K strata. Consequently, it is necessary to add a second subscript, $k = 1, 2, \cdots, K$, everywhere, except that $\mu_{1k}/\mu_{2k} = \delta_L$ is assumed constant for all k. We now consider testing the null hypothesis $H_{0K}^1 : \mu_{Ek}/\mu_{Rk} \le \delta_L$ for all $k = 1, 2, \cdots, K$ against the alternative hypothesis $H_{AK}^1 : \mu_{Ek}/\mu_{Rk} \ge \delta_L$ for all k, and $\mu_{Ek}/\mu_{Rk} > \delta_L$ for some k. If we choose the simple null hypothesis

$$H_{0K}^1 : \mu_{E1}/\mu_{R1} = \mu_{E2}/\mu_{R1} = \cdots = \mu_{EK}/\mu_{RK} = \delta_L,$$

to be the one containing the equality statement, we have

$$\mu_{E1}/\mu_{R1} = \mu_{E2}/\mu_{R1} = \cdots = \mu_{EK}/\mu_{RK} = \sum_1^K \mu_{Ek} \Big/ \sum_1^K \mu_{Rk} = \delta_L.$$

That is, it is possible to restate the null and alternative hypotheses in terms of the sums of strata medians. Let $\mu_1 = \sum_1^K \mu_{Ek}, \mu_2 = \sum_1^K \mu_{Rk}$. Then our objective is to test

$$H_{0K}^1 : \mu_1/\mu_2 \le \delta_L \text{ vs. } H_{AK}^1 : \mu_1/\mu_2 > \delta_L \qquad (9.1)$$

Consequently, we set $X_1 = \sum_1^K m_{Ek}, X_2 = \sum_1^K m_{Rk}$, $v_1^2 = \sum_1^K \sigma_{Ek}^2$ and $v_2^2 = \sum_1^K \sigma_{Rk}^2$. Now X_1 is normally distributed with mean μ_1 and variance v_1^2 and X_2 is

Table 1. Optimal numbers of events *r per arm for $\alpha = 0.025$ and $d = 0.45$.**

λ_E	λ_R (median)					
	0.01 (69.3)	0.02 (34.7)	0.025 (27.7)	0.03 (23.1)	0.04 (17.33)	0.05 (13.86)
0.01	158					
0.02		158	243			
0.025		103	158	225		
0.03			112	158	276	
0.04				92	158	243
0.05					103	158

Table 2. Optimal numbers of events *r per arm for $\alpha = 0.05$ and $d = 0.45$.**

λ_E	λ_R (median)					
	0.01 (69.3)	0.02 (34.7)	0.025 (27.7)	0.03 (23.1)	0.04 (17.33)	0.05 (13.86)
0.01	111					
0.02		111	172			
0.025		73	111	158		
0.03			73	111	195	
0.04				65	111	172
0.05					73	111

normally distributed with mean μ_2 and variance v_2^2. Now let $W = X_1/X_2$. The ratio W follows the F-H distribution. The null distribution function of W is given by

$$G\left(w\middle|H_{0K}\right) = \Phi\left[\frac{\mu_2\left(w-\delta_L\right)}{\hat{v}_1\hat{v}_2 a\left(w\right)}\right], \delta_L = \mu_1/\mu_2 \text{, and}$$

$$a\left(w\right) = \left(\frac{w^2}{\hat{v}_1^2} + \frac{1}{\hat{v}_2^2}\right)^{1/2}, \tag{9.2}$$

where $\hat{v}_1^2$ and $\hat{v}_2^2$ are estimates of v_1^2 and v_2^2, respectively. Reject $H_{0K}^1 : \mu_1/\mu_2 \le \delta_L$ in favor of $H_{AK}^1 : \mu_1/\mu_2 > \delta_L$ if $W > w'$, where $G\left(w'\middle|H_{0K}^1\right) = 1-\alpha$. The cut-off point w' satisfies the equation $\mu_2\left(w'-\delta_L\right) = z_\alpha \hat{v}_1\hat{v}_2 a\left(w'\right)$. Note that $w' = w'\left(\mu_2\right)$.

Let $\ell = \left\{\left(\mu_1,\mu_2\right)' : \mu_1 = \delta_L \times \mu_2\right\}$, where Θ_ω is a rectangle defined by the $100\left(1-\omega\right)$ percent confidence intervals on μ_1 and μ_2. As earlier, let $\tilde{\mu}_2$ represent the smallest μ_2 in ℓ and $\tilde{w} = w'\left(\tilde{\mu}_2\right)$. This results in $G\left(\tilde{w}\middle|H_{0K}^1\right) \le \alpha$. Therefore, the rule that rejects H_{0K}^1 in favor of H_{AK}^1 for $W > \tilde{w}$ is a level α test.

10. Test for Equivalence

The objective is to test

$$H_0 : \frac{\mu_E}{\mu_R} \le \delta_L \text{ or } \frac{\mu_E}{\mu_R} \ge \delta_U \text{ vs } H_A : \delta_L < \frac{\mu_E}{\mu_R} < \delta_U \text{ , (10.1)}$$

where the interval $\left(\delta_L,\delta_U\right)$ is called equivalence range in clinical trials terminology. The equivalence range may be of the form $\left(\delta^{-1},\delta\right)$ for some $\delta > 1$. We use the well-known two one-sided tests (TOST) approach to test the null hypothesis H_0 against the alternative hypothesis H_A given in (10.1). We first test the following two one-sided hypotheses

$H_0^1 : \mu_E/\mu_R \le \delta_L$ vs $H_A^1 : \mu_E/\mu_R > \delta_L$, and

$H_0^2 : \mu_E/\mu_R \le \delta_U$ vs $H_A^2 : \mu_E/\mu_R < \delta_U$

and then combine the results according intersection-union principle. We have already outlined the two one-sided tests in Sections 5 and 6 above. The null hypothesis H_0 is rejected in favor of H_A at level α, if both hypotheses H_0^1 and H_0^2 are rejected at level α. As indicated by Berger and Hsu [9], this test can be quite conservative. We define the *p*-value as the $\max\left(\alpha_L', \alpha_U'\right)$, where α_L' and α_U' are defined in (5.5) and (6.3), respectively.

Next, we discuss the power of the test of H_0 versus H_A of (10.1). We evaluate the power of the test at the alternative $\mu_E/\mu_R = 1$. Note that we reject H_0^1 if $W > \tilde{w}_1$ and we reject H_0^2 if $W < \tilde{w}_2$, where $\tilde{w}_1$ and $\tilde{w}_2$ are determined as explained in Sections 5 and 6, respectively. Intuitively, the power of the test is

$$\beta = 1 - P\left(\tilde{w}_2 < W < \tilde{w}_1 \middle| \frac{\mu_E}{\mu_R} = 1\right), \tilde{w}_2 < \tilde{w}_1$$

For $\tilde{w}_2 < \tilde{w}_1$, the power $\beta = \beta(1)$ is

$$\beta = \beta(1)$$
$$= 1 - \left(\Phi\left[\frac{\tilde{\mu}_{R1}\left(\tilde{w}_1 - 1\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(\tilde{w}_1\right)}\right] - \Phi\left[\frac{\tilde{\mu}_{R2}\left(\tilde{w}_2 - 1\right)}{\hat{\sigma}_1\hat{\sigma}_2 a\left(\tilde{w}_2\right)}\right]\right). \text{ (10.2)}$$

However, this power may be low in some cases. Then one may use **Table 1** or **Table 2** for sample size determination.

In **Figure 2** we have provided a graphical summarization of testing for equivalence at $\alpha = 0.05$.

Figure 2 contains the density functions of W for non-inferiority margin $\delta = 0.8, 1.0, 1.25$. Here we have used $\hat{\sigma}_1 = 1.42$, and $\hat{\sigma}_2 = 2.787$ in all three cases. Note that $\tilde{w}_1 = 1.51$ and $\tilde{w}_2 = 0.67$ are the cutoff points and the area marked by (1) and (2) represent the level of significance α for testing H_0^1 and H_0^2, respectively. The total area represented by (1) + (2) + (3) + (4) is the power of the equivalence test given in (10.2).

11. Concluding Remarks

We deal with the ratio μ_E/μ_R directly, and therefore,

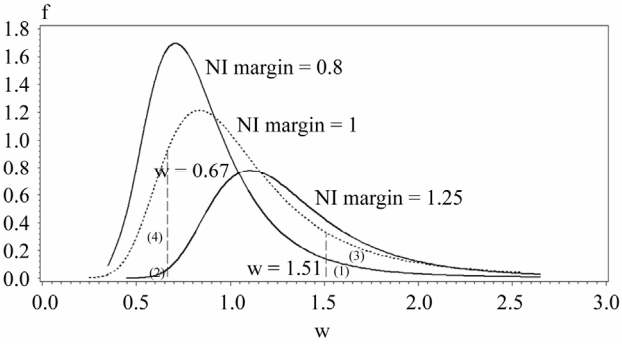

Figure 2. Overview of the equivalence test.

our approach is easy for clinicians to understand. Existing test procedures for assessing non-inferiority and equivalence require hazard rates under the two treatment arms to be proportional. Our test proposed in this paper is free of this requirement and therefore, has wider applicability.

The power definitions in (5.6) and (6.4) may be considered as alternative to the power definitions in [20,21].

It may be recalled here that the Mantel-Haenszel test [23] is often called an *average* partial association statistic. Here we have a parallel situation. Note that the null hypothesis H_{0K}^1 in (9.1) may be written as $H_{0K}^1 : \bar{\mu}_E / \bar{\mu}_R \le \delta_L$, where $\bar{\mu}_E = \sum_1^K \mu_{Ek} / K$, and $\bar{\mu}_R = \sum_1^K \mu_{Rk} / K$. Therefore, the procedure in Section 9 tests the null hypothesis on the ratio of averages of strata medians.

12. Acknowledgements

This article reflects the views of the author and should not be construed to represent FDA's views or policies. No official support or endorsement of this article by the Food and Drug Administration is intended or should be inferred.

REFERENCES

[1] "E-10: Guidance on Choice of Control Group in Clinical Trials," *International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use* (*ICH*), Vol. 64, No. 185, 2000, pp. 51767-51780.

[2] R. B. D'Agostino, J. M. Massaro and L. M. Sullivan, "Non-Inferiority Trials: Design Concepts and Issues—The Encounters of Academic Consultants in Statistics," *Statistics in Medicine*, Vol. 22, No. 2, 2003, pp. 169-186. doi:10.1002/sim.1425

[3] G. G. Koch, "Non-Inferiority in Confirmatory Active Control Clinical Trials: Concepts and Statistical Methods," American Statistical Association: FDA/Industry Workshop, Washington, D.C., 2004.

[4] S. Wellek, "Testing Statistical Hypothesis of Equivalence," CHAPMAN & HALL/CRC, New York, 2003.

[5] B. Efron, "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, Vol. 76, No. 374, 1981, pp. 312-319. doi:10.1080/01621459.1981.10477650

[6] R. Simon, "Confidence Intervals for Reporting Results of Clinical Trials," *Annals of Internal Medicine*, Vol. 105, No. 3, 1986, pp. 429-435.

[7] L. Rubinstein, M. Gail and T. Santner, "Planning the Duration of a Comparative Clinical Trial with Loss to Follow-Up and a Period of Continued Observation," *Journal of Chronic Disease*, Vol. 34, No. 9-10, 1981, pp. 469-479. doi:10.1016/0021-9681(81)90007-2

[8] D. R. Bristol, "Planning Survival Studies to Compare a Treatment to an Active Control," *Journal of Biopharmaceutical Statistics*, Vol. 3, No. 2, 1993, pp. 153-158. doi:10.1080/10543409308835056

[9] R. L. Berger and J. C. Hsu, "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," *Statistical Science*, Vol. 11, No. 4, 1996, pp. 283-319. doi:10.1214/ss/1032280304

[10] D. Hauschke and L. A. Hothorn, "Letter to the Editor," *Statistics in Medicine*, Vol. 26, No. 1, 2007, pp. 230-236. doi:10.1002/sim.2665

[11] SAS Institute Inc., "SAS/STAT User's Guide," Version 8, Cary, 2000.

[12] R. Brookmeyer and J. Crowley, "A Confidence Interval for the Median Survival Time," *Biometrics*, Vol. 38, No. 1, 1982, pp. 29-41. doi:10.2307/2530286

[13] D. Collett, "Modeling Survival Data in Medical Research," 1st Edition, Chapman & Hall, London, 1994.

[14] N. Reid, "Estimating the Median Survival Time," *Biometrika*, Vol. 68, No. 3, 1981, pp. 601-608. doi:10.1093/biomet/68.3.601

[15] G. J. Babu, "A Note on Bootstrapping the Variance of Sample Quantiles," *Annals of the Institute of Statistical Mathematics*, Vol. 38, 1985, pp. 439-443. doi:10.1007/BF02482530

[16] The University of Texas at Austin, "Setting and Resampling in SAS," 1996. http://ftp.sas.com/techsup/download/stat/jackboot.htm/

[17] K. M. Keaney and L. J. Wei, "Interim Analyses Based on Median Survival Times," *Biometrika*, Vol. 81, No. 2, 1994, pp. 279-286. doi:10.1093/biomet/81.2.279

[18] E. C. Fieller, "The Distribution of the Index in a Normal Bivariate Population," *Biometrika*, Vol. 24, No. 3-4, 1932, pp. 428-440. doi:10.1093/biomet/24.3-4.428

[19] D. V. Hinkley, "On the Ratio of Two Correlated Normal Variables," *Biometrika*, Vol. 56, No. 3, 1969, pp. 635-639. doi:10.1093/biomet/56.3.635

[20] K. M. Koti, "Use of the Fieller-Hinkley Distribution of the Ratio of Random Variables in Testing for Non-Inferiority and Equivalence," *Journal of Biopharmaceutical Statistics*, Vol. 17, No. 2, 2007, pp. 215-228. doi:10.1080/10543400601177335

[21] K. M. Koti, "New Tests for Null Hypothesis of Non-Unity Ratio of Proportions," *Journal of Biopharmaceuti-*

cal Statistics, Vol. 17, No. 2, 2007, pp. 229-245. doi:10.1080/10543400601177426

[22] B. Efron and R. J. Tibshirani, "An Introduction to the Bootstrap," Chapman & Hall, New York, 1993.

[23] M. E. Stokes, C. S. Davis and G. G. Koch, "Categorical Data Analysis Using the SAS System," SAS Institute Inc., Cary, 1995.

