行政院國家科學委員會專題研究計畫　成果報告

二存活函數之對等性檢定

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-004-002-
執行期間：91 年 08 月 01 日至 93 年 09 月 30 日
執行單位：國立政治大學統計學系

計畫主持人：薛慧敏

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 93 年 9 月 21 日

Sample Size for Evaulaiton of Equivalence and Non-inferiority Tests in the Comparison of Two Survival Functions

H. M. Hsueh,[1] **J. P. Liu**[2] **and T. J. Yao**[2,*]

[1] Department of Statistics, National Cheng Chi University, Taipei, Taiwan

[2]Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan

Summary. In oncology, increasing number of active control trials have been conducted to compare a test therapy to a standard therapy. These new therapies are developed for less invasive or easy administration, or for reduced toxicity and thus to improve the quality of life at the minimal expense of survival. Therefore, evaluation of equivalence or non-inferiority based on censored endpoints such as overall survivals between test and active control becomes an important and practical issue. Under the assumption of proportional hazards, Wellek (1993) proposed a log-rank test for assessment of equivalence of two survival functions. In this paper, an explicit form of the asymptotic variance of the maximum likelihood estimator for the treatment effect is derived. It follows that the asymptotic power and sample size formulae can also be obtained. Alternatively, a two one-sided test (TOST) is proposed to evaluate the equivalence of two survival functions. The critical values of the proposed TOST depend upon only the asymptotic variance

* *email:* tjyao@nhri.org.tw

and the standard normal percentiles, which greatly simplify the sample size determination. In addition, a procedure for testing non-inferiority based on censored endpoint is derived and the corresponding sample size formula is also provided. It can be shown that when the sample size is large, the same sample size formulae can be derived for both the log-rank test and TOST when two survival functions are assumed to be equal. The sample size formulas for both procedures take into account the accrual pattern and the duration of the study. A simulation is conducted to empirically investigate the performance on size, power, and sample size of the proposed procedures and the log-rank test. Numerical examples are provided to illustrate the proposed procedures.

KEY WORDS: Equivalence, Non-inferiority, Survival Function, Two one-sided test procedure, Power, Sample size

1. Introduction

There are increasing number of comparative clinical trials where new treatments are compared to an active control. Many active control trials are designed due to ethical concerns, especially for life-threatening diseases. However, for many others, the new treatments are developed for less invasive or easier administration, or for reduced toxicity and thus to improve the quality of life at minimal cost of efficacy. Therefore, the assessment of equivalence or non-inferiority in efficacy is an important and practical issue. For example, investigators in gynecological oncology were interested in the evaluation of the efficacy and toxicity of adjuvant therapy after radical hysterectomy and pelvic lymphadenectomy in stage IB or IIA cervical carcinoma patients with pelvic lymph node metastases. The current standard adjuvant therapy is the concomitant chemo-radiotherapy (CT+RT) (cisplatin plus radiotherapy). However, postoperative radiotherapy is associated with significant morbidity. Some retrospective studies indicated that the adjuvant chemotherapy alone (CT) (cisplatin, onvovin and bleomycin) seemed to have comparable recurrenc-free survival rate and significantly less morbidity. The investigators would like to confirm the findings with a randomized trial and prove that the recurrence-free survival for patients treated with CT and for patients treated with CT+RT are equivalent.

For another example, axillary lymph nodes dissection has been a standard procedure in patients with breast carcinoma for staging and the prevention of metastases. However, the disease for the majority of early stage patients have not metastasized to the lymph nodes and this procedure can cause complications in the axillary area and the upper arm. To avoid the unnecessary complications, investigators have developed techniques in the identification of the sentinel nodes. It is then of interest to see if the disease-free survival of patients with sentinel nodes dissection alone is as good as that of patients with axillary nodes dissection.

For patients with stage IV nasopharyngeal carcinoma (NPC), concurrent chemoradiotherapy (cisplatin/5-FU) plus adjuvant (cisplatin/5-FU) has been proved to prolong survival. However, the adjuvant therapy may increase the chances of long term toxicity. At least 20% patients suffered severe toxicity, mostly vomiting or mucositis, while receiving adjuvant chemotherapy, and more than 11% patients surffered late complication, mostly hearing impairment or neck fibrosis (Cheng et al., 2000). Certain herbal medicine were then proposed to replace the adjuvant therapy in the hope to maintain the survival effect but reduce the risk of side effects. It is thus relevant to have a comparative trial to test the hypothesis that five-year survival rate for this herbal medicine is equivalent to that afforded by cisplatin/5-FU in this patient population.

Several works have been published to address the issues in design and analysis of equivalence or non-inferiority trials for binary endpoints, such as Hsueh, Liu and Chen (2001), Kang and Chen (2000), Chan (1998), and Farrington and Manning (1990), among others. Much less work has been developed when the primary endpoint is the time-to-failure data. However, many equivalence/non-inferiority trials are designed for treatments to life-treatening disease, and for which time-to-failure is usually the primary endpoint, as described in the previous three examples. Literature for the equivalence in time-to-failure data is scarce. Under proportional hazards assumptions, Wellek (1993) proposed a log-rank test based on the asymptotic normality of the maximum partial likelihood estimator and an approximation form of the asymptotic variance. The test also used a noncentral chi-square percentile with a data-dependent noncentrality parameter as the critical value.

In this paper, we follow the development of Wellek and further derive an explicit form of the asymptotic variance. With this form, sample size determination is possible without simulation studies. In addition, we proposed the two one-sided test (TOST) as an alternative of the log-rank test. The critical values of TOST only depend on the asymptotic variance and the standard normal percentiles, which greatly simplifies sample size determination. We also allocate much effort in pointing out issues in the application of the results to actual trial designs. In Section 2, some basic notations and Wellek's test are introduced. The asymptotic variance is derived explicitly. To reduce the complexity in the calculation of the critical value, an alternative two one-sided test(TOST) is proposed in Section 3. Corresponding sample size formulas are given in Section 4. Some simulation and numerical results are presented in Section 5 and an example is illustrated in Section 6. Discussion and final remarks are provided in Section 7.

2. The asymptotic variance

We consider that two unrelated samples $(T_1, \cdots, T_{n_1})$, and $(T_{n_1+1}, \cdots, T_{n_1+n_2})$ of possibly right censored survival times are given such that $T_i \sim S_1(\cdot)$ for $1 \le i \le n_1$ and $T_i \sim S_2(\cdot)$ for $n_1 + 1 \le i \le n_1 + n_2$. We assume that $S_1(\cdot)$ and $S_2(\cdot)$ belong to the same proportional hazards model and with no loss of generality, $S_1(\cdot)$ is the survivor function for the control group. That is,

$$S_2(t) = \{S_1(t)\}^{e^\theta} \quad \text{for all } t > 0 \text{ and some } \theta.$$

Hence, for $i = 1, \cdots, n_1 + n_2$, T_i has hazard function

$$\lambda(t) = \lambda_1(t)e^{z_i\theta}$$

where

$$\lambda_1(t) = \frac{d}{dt}\{-\log S_1(t)\}, \quad z_i = \begin{cases} 0, & \text{for } 1 \le i \le n_1; \\ 1, & \text{for } n_1 + 1 \le i \le n_1 + n_2. \end{cases}$$

We considered the hypotheses of equivalence by restricting the uniform absolute difference between $S_1(\cdot)$ and $S_2(\cdot)$ as the following

$$H_0 : \sup_{t>0} |S_1(t) - S_2(t)| \ge \delta \text{ versus } H_1 : \sup_{t>0} |S_1(t) - S_2(t)| < \delta, \qquad (1)$$

for some $\delta > 0$. Wellek (1993) showed that, by the continuity of $S_1(\cdot)$ and reparametrization, the hypotheses (1) are equivalent to

$$H_0^a : |\theta| \geq \theta^* \text{ versus } H_1^a : |\theta| < \theta^* \tag{2}$$

where θ^* satisfies

$$e^{\theta^*/(1-e^{\theta^*})} - e^{\theta^* e^{\theta^*}/(1-e^{\theta^*})} = \delta.$$

The partial log-likelihood for θ can be shown as

$$\log L(\theta) = d_{+2}\theta - \sum_{j=1}^{k} d_{j+} \log\{r_{j1} + r_{j2}e^{\theta}\},$$

where for $v = 1, 2; j = 1, \cdots, k$,

r_{jv} = number of items at risk in the vth sample at the jth smallest failure time $t_{(j)}$,

d_{+v} = total number of failures in the vth sample,

d_{j+} = total number of failures at the jth smallest failure time.

Hence, the maximum partial likelihood estimator $\hat{\theta}$ satisifies

$$\sum_{j=1}^{k} d_{j+} \frac{r_{j2}e^{\hat{\theta}}}{(r_{j1} + r_{j2}e^{\hat{\theta}})} = d_{+2},$$

and the observed information at $\hat{\theta}$ is

$$I(\hat{\theta}) = \sum_{j=1}^{k} \frac{d_{j+} r_{j1} r_{j2} e^{\hat{\theta}}}{(r_{j1} + r_{j2}e^{\hat{\theta}})^2},$$

as $N = n_1 + n_2 \to \infty$

$$I(\hat{\theta})/N \xrightarrow{p} 1/v^2(\theta),$$

where the reciprocal $1/v^2(\theta)$ is the limiting value of the observed information and is a function of true value θ.

Defining $\tilde{C}_\alpha(\psi)$ as

$$\tilde{C}_\alpha(\psi) \;=\; \{\ \alpha\text{th quantile of a } \chi^2 \text{ distribution with df=1 and noncentrality parameter } \psi^2\ \}^{1/2}$$

for arbitrary $\psi > 0$, Wellek (1993) proposed an aymptotic UMP level α test with rejection region

$$\sqrt{N}|\hat{\theta}|/v(\hat{\theta}) < \tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\hat{\theta})\right\}. \tag{3}$$

Without an explicit form of $v(\theta)$, Wellek suggested estimating $\sqrt{N}/v(\hat{\theta})$ by the observed information $\sqrt{I(\hat{\theta})}$. This test procedure is later referred as the log-rank test for equivalence in this paper.

In fact, an explicit form of the limiting value $v^2(\theta)$ can be derived. Let T be the event time, C be the censoring time and Δ be the censoring indicator, i.e. $\Delta = 1$ if $T \leq C$ and $\Delta = 0$ if $T > C$. We define f_i as the p.d.f. corresponding to the survival function S_i, and $S_{ci}(t) = Pr(C > t|Z = i - 1)$ as the censoring function for group i, $i = 1, 2$. In addition, let $\rho = \lim_{N\to\infty} n_2/N$, and $0 < \rho < 1$. Therefore, we have the following results.

THEOREM 1. *If the trial has an infinity duration, then*

$$1/v^2(\theta) = \int_0^\infty p(s)q(s)u(s)ds.$$

where

$$\begin{aligned} p(s) &= Pr(Z = 1 \mid T = s, \Delta = 1) = \frac{\rho S_{c2}(s)f_2(s)}{\rho S_{c2}(s)f_2(s) + (1-\rho)S_{c1}f_1(s)} \\ &\equiv 1 - q(s), \\ u(s) &= Pr(T = s, \Delta = 1) = \rho S_{c2}(s)f_2(s) + (1-\rho)S_{c1}f_1(s). \end{aligned}$$

The proof of Theorem 1 is given in the Appendix. With this explicit form of $v^2(\theta)$, the asymptotic behavior of the log-rank test can be evaluated analytically, and the sample size determination is possible without simulation, as will be shown in Section 4. However, in practice, a trial is always designed within a finite period of accrual time plus an additional finite period of follow-up time. Therefore, the following result is much more relevant and useful.

Corollary 1. *If the accural period of the trial is T_0, and an additional follow-up period of τ is considered, assuming an uniform accrual rate, then*

$$1/v^2(\theta) = \int_0^{T_0} \int_0^{T_0+\tau-t} p(s)q(s)u(s)ds \frac{1}{T_0} dt.$$

3. A two one-sided test (TOST)

Wellek's test procedure is complicated to apply because of the necessity to evaluate the noncentral chi-square percentile in (3) . It is easy to see that the hypotheses in (2) can be partitioned into two one-sided hypotheses,

$$H_{U0}^a : \theta \leq -\theta^* \text{ versus } H_{U1}^a : \theta > -\theta^*$$

and

$$H_{L0}^a : \theta \geq \theta^* \text{ versus } H_{L1}^a : \theta < \theta^*.$$

By the intersection-union principle, H_0^a is rejected if both H_{U0}^a, H_{L0}^a are rejected. Thus the rejection region of the two one-sided test procedure (TOST) at level α can be easily shown as

$$Z_L = (\hat{\theta} - \theta^*)\sqrt{I(\hat{\theta})} < -z_\alpha, \text{ and } Z_U = (\hat{\theta} + \theta^*)\sqrt{I(\hat{\theta})} > z_\alpha. \qquad (4)$$

Denotes $\Phi(\cdot)$ as the standard normal distribution function. The following theorem gives the asymptotic power functions of the log-rank test (3) and the two one-sided test (4), respectively.

Theorem 2. *At $\theta = \theta^0, |\theta^0| < \theta^*$, the log-rank test (3) has asymptotic power*

$$\begin{aligned}\beta_{LR}(\theta^0) &= \Phi\left(\tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right\} - \sqrt{N}\theta^0/v(\theta^0)\right) - \\ &\quad \Phi\left(-\tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right\} - \sqrt{N}\theta^0/v(\theta^0)\right);\end{aligned}$$

and the two one-sided test (4) has asymptotic power

$$\begin{aligned}\beta_{TOST}(\theta^0) &= \Phi\left(-z_\alpha + \sqrt{N}\theta^*/v(\theta^0) - \sqrt{N}\theta^0/v(\theta^0)\right) - \\ &\quad \Phi\left(z_\alpha - \sqrt{N}\theta^*/v(\theta^0) - \sqrt{N}\theta^0/v(\theta^0)\right).\end{aligned}$$

Furthermore, the TOST test for the hypotheses of equality can be easily modified to a test for the hypotheses of non-inferiority:

$$H_{0L} : \inf_{t>0}\{S_2(t) - S_1(t)\} \leq -\delta \text{ versus } H_{1L} : \inf_{t>0}\{S_2(t) - S_1(t)\} > -\delta,$$

or equivalently,

$$H^a_{0L} : \theta \geq \theta^* \text{ versus } H^a_{1L} : \theta < \theta^*.$$

The corresponding one-sided test procedure based on Z_L in (4) is used.

Corollary 2. *At $\theta = \theta^0, \theta^0 < \theta^*$, the non-inferiority test procedure has asymptotic power*

$$\beta_{NI}(\theta^0) = \Phi\left(-z_\alpha + \sqrt{N}\theta^*/v(\theta^0) - \sqrt{N}\theta^0/v(\theta^0)\right).$$

4. Sample size determination

At the design stage of clinical trials, determination of the sample size is always a key element. It plays an important role in assessing the feasibility of a trial. For clinical trials to test the equivalence between the survival functions of two study arms, the sample size is often required to acheieve a pre-determined level of power, β^*, at $S_1(\cdot) = S_2(\cdot)$, or $\theta^0 = 0$ in proportional hazards models. With the explicit form of $v(\theta)$, the sample size formulae of the log-rank test, TOST and the non-inferiority test can be derived analytically.

Corollary 3. *For the asymptotic power greater than β^* at $\theta^0 = 0$, the sample size required for the log-rank test (3) is the smallest integer N such that*

$$\tilde{C}_\alpha \left\{ \sqrt{N}\theta^* / v(0) \right\} \geq z_{(1-\beta^*)/2}. \tag{5}$$

The TOST (4) should have sample size no less than

$$\left\{ z_\alpha + z_{(1-\beta^*)/2} \right\}^2 v^2(0)/\theta^{*2},$$

while the sample size required for the non-inferiority test Z_L is at least

$$\left\{ z_\alpha + z_{(1-\beta^*)} \right\}^2 v^2(0)/\theta^{*2}.$$

Notice that, at $\theta^0 = 0$, $p(s), q(s)$ and $u(s)$ in $v^2(0)$ are simplified as

$$p(s) = \frac{\rho S_{C2}(s)}{\rho S_{C2}(s) + (1-\rho) S_{C1}(s)} = 1 - q(s),$$

and

$$u(s) = f_1(s)\{\rho S_{C2}(s) + (1-\rho) S_{C1}(s)\},$$

respectively. We can see that the sample size evaluation for TOST only requires one calculation involving standard normal percentiles, while it requires repeated calculations of noncentral chi-square percentiles for the log-rank test.

Even though the preceding test procedures can be applied to any proportional hazards models without specific forms of the survivor functions, the determination of the sample size depends on the specified $f_1(\cdot), S_{C1}(\cdot)$ and $S_{C2}(\cdot)$ in $v^2(0)$.

5. Simulation results

Several simulation studies were performed in order to study the small sample performance of the three test procedures. The first simulation study used the same parameters as the one reported in Wellek for comparison. The samples of the control arm were generated under a lognormal survivor function, $S_1(t) = \Phi(2 - \ln(t))$ and an independent exponential censoring distribution function, $S_{C1}(t) = 1 - exp\{-t/50\}$. The censoring rate for this arm is about 19%. The samples of the treatment arm were generated at both boundaries of the hypothesis with $\delta = 0.15$ and at an identical alternative $S_2 = S_1$, all with $S_{C2} = S_{C1}$. The empirical type I error rates and power were thus calculated based on 10,000 replications. The computations were carried out using FORTRAN 90 on an x86 Family 6 Model 8 Stepping 10 PC. The results of different sample sizes are presented in Table 1-1. The corresponding approximation of the asymptotic size and power were also evaluated at $\theta^0 = \pm\theta^*$ and $\theta^0 = 0$, respectively, and presented in the same Table. Table 1-2 shows the simulation results under the same distributional assumptions as those for Table 1-1 except that the maximum allowable difference δ is reduced to

0.10. For the other 2 simulation studies, the survivor function for the control group was specified in an exponential model: $S_1(t) = exp(-t/\lambda_1)$, where λ_1 is selected such that $S_1(5) = 0.55$, and the censoring distribution was $S_{C1}(t) = S_{C2} = exp(-t/4\lambda_1)$. Note that the censoring distribution was selected such that the censoring rate is 20% for infinite duration. The results are shown in Table 2-1 when $\delta = 0.15$ and Table 2-2 when $\delta = 0.10$.

In general, TOST is more conservative when compared to the log-rank test. For very small sample sizes, TOST is extremely conservative, but the differences between TOST and the log-rank test decreases as the sample size increases. Under both distribution models, when $\delta = 0.15$, the type I error rate and power for the log-rank test and TOST are virtually identical after the total sample size reaches 250, where the type I error rate is approximately controlled. In fact, it can be shown that the sample sizes for both tests approximate to the same value for large N. Let χ^2_ψ be a chi-square random variable with noncentrality parameter ψ^2, then

$$\alpha = P(\chi^2_\psi \leq \tilde{C}^2_\alpha(\psi)) = \Phi(\tilde{C}_\alpha(\psi) - \psi) - \Phi(-\tilde{C}_\alpha(\psi) - \psi).$$

We see that as ψ goes to infinity,

$$\Phi(\tilde{C}_\alpha(\psi) - \psi) \approx \alpha, \text{ and } \tilde{C}_\alpha(\psi) \approx \psi - z_\alpha.$$

That is, the asymptotic power of the log-rank test $\beta_{LR}(\theta^0)$ can be approximated by the asymptotic power of TOST $\beta_{TOST}(\theta^0)$ provided N is sufficiently large. If a trial to test the equivalence of two survivor functions is designed with sufficient power, say the most commonly required 80%, the sufficient total sample size will be greater than 250. In other words, with reasonable sample sizes, TOST has the same performance as the log-rank test and has the advantage of easily evaluable rejection region and power.

We can see that the maximum allowable difference δ is very influential on sample size determination. Under both survivor models, when δ reduces to 0.10, the performance of the log-rank test and TOST becomes identical as the total sample size reaches 500. For a test with 80% power, the required sample size is close to 600.

6. Numerical examples

In the example of stage IV NPC, the 5-year survival rate has been reported to be 55% when treated with CCRT plus adjuvant chemotherapy (Cheng et al, 2000). Medical investigators are interested to see if the survivor function remain unchanged with the substitution of the adjuvant chemotherapy by a herbal medicine after CCRT. A randomized clinical trial is then designed to test the hypothesis of equivalence in the survivor functions of the two arms. Table 3 shows the sample sizes for the log-rank test and TOST under a log-normal model or an exponential model. The parameter for either model was determined so that $S_1(5) = 0.55$. In addition, exponential models are used for the censoring distributions, for which the parameters were determined so that the censoring rate was 20% under either survival model. That is, for the log-normal model, $S_1(t) = \Phi(1.735 - ln(t))$ and $S_{C1}(t) = S_{C2}(t) = 1 - exp\{-t/35.7\}$; and for the exponential model,$S_1(t) = 1 - exp\{-t/8.36\}$ and $S_{C1}(t) = S_{C2}(t) = 1 - exp\{-t/33.5\}$. Equal allocation for the two arms was assumed, and δ was set at 0.15. For illustration, the sample sizes for the non-inferiority test are also shown in Table 3. We note that under the case of $S_{C1}(t) = S_{C2}(t)$, and at $\theta = 0$, Δ and T are independent with Z. For $\forall s$,

$$p(s) = P(Z = 1|T, \Delta) = P(Z = 1) = \rho = 1 - q(s),$$

then

$$1/v^2(0) = \rho(1-\rho)\int_0^\infty P(T = s, \Delta = 1)ds = \rho(1-\rho)P(\Delta = 1)$$

is proportional to the probability of uncensored observations regardless of the specifications of S_1 and S_{C1}. Therefore, the sample size is inversely proportional to the uncensoring proportion. In our example, since the censoring rate and the allocation proportion are the same for both models, they have the identical sample sizes for trials with infinite duration.

In addition to the sample size calculations for trials with infinite duration, we also calculated the sample sizes for the more realistic cases with 5 years of uniform accrual and 1 or 2 additional years of follow-up, using the expression for $1/v^2(\theta)$ in Corollary 1. The results indicate that when the accrual duration is limited to 5 years, the sample size is approximately doubled or tripled for additional 2 or 1 year of follow-up. With these specifications, sample sizes under the log-normal model are larger than those under exponential model by about 10% for the cases with 1 year follow-up and 5% for the cases with 2 years of follow-up. Similar to the unlimited case, with independence between Δ, T and Z at $\theta = 0$, the values of $1/v^2(\theta)$ for the limited accrual case is proportional to

$$E^U E^T\{E(\Delta|T)|T < T_0 + \tau - U\},$$

and hence the sample size for the limited accrual case is inversely proportional to the probability of uncensored observation up to time $T_0 + \tau$. Thus the difference in sample size between the unlimited and the limited accrual cases is believed to increase as the uncensoring proportion within the limited study period decreases. We like to emphasize the fact that the sample sizes for the log-rank test and for TOST are identical for all cases in Table 3.

7. Discussion

We have proposed a two one-sided test for testing the equivalence of two survivor curves. With moderate or large sample sizes, the type I error rates evaluated at the boundaries and the power evaluated at $S_1 = S_2$ for TOST are virtually identical to those for the log-rank test proposed by Wellek. These tests are developed under the assumption of proportional hazards. However, as in most of the designs for comparative clinical trials, distributional models need to be specified for sample size determination. We like to point out that, when we reject the null hypothesis in a log-rank test or TOST, the maximal difference between $S_1(t)$ and $S_2(t)$ is equal to or smaller than an acceptable boundary δ. It is easy to show that for all proportional hazards models, this maximum occurs at time t^* where $S_1(t^*) = exp\,[\theta^*/\{1 - exp(\theta^*)\}]$ and θ^* is the corresponding equivalence limit. Therefore, the sample size required to perform these tests with certain power could be very different from the the sample size required to test the equivalence of survival at a fixed time.

For example, the 5-year disease-free survival is about $S_1(5) = 95\%$ for patients with early stage breast carcinoma after surgery with axillary lymph nodes dissection. A comparative trial is designed to test the non-inferiority in disease-free survival for patients with sentinel nodes dissection (STND) instead of axillary nodes dissection (AXND). The investigators consider STND group as non-inferior if the 5-year disease-free survival $S_2(5)$ is at least 90%, that is , $\delta = 5\%$, then under exponential models and $S_2(t) = S_1(t)^{exp(\theta)}$, $\theta = 0.7198$ at the boundary $S_2(5) = 90\%$, which leads to a very small sample size, 24 per arm, for equal allocation and infinite accrual without censoring. We note here that if we relax the model assumption and calculate the sample size to test the non-inferiority between two proportions using method porposed by Kang and Chen (2000), then we need 251 patients per arm and each patient has to be followed for 5 years.

However, if the investigators consider STND to be non-inferior to AXND only if the difference between survival is at most 5% at any time, then it is the same as requiring $S_2(t^*) \geq S_1(t^*) - 5\%$ at time t^*, where t^* is solved to be 90.5 years! This leads to $\theta^* = 0.1360$ and a sample size of 669 per arm. Neither of the two results seem appealing. The former case concentrates on one single time point and thus highly depends on the model assumptions. The latter case considers the time period that goes beyond any practically meaningful length. One reasonable alternative is to let investigators determine the lengh of interest, say 15 years, and then test the non-inferiority within this time period. Since $|S_1(t) - S_2(t)|$ is an unimodal function of t and $15 < t^*$, it is easy to see that the maximum for this period $[0, 15]$ occurs at $t^0 = 15$. This leads to $\theta^0 = 0.3297$ and a sample size of 114 per arm.

In general, the sample sizes sufficient to prove equivalence or non-inferiority are larger than the sample sizes sufficient to prove difference. In designing trials for equivalence or non-inferiority based on censored endpoints, there are several non-trivial issues as pointed out in this paper. The investigators should be aware of the implications of the assumptions in sample size caculation in order to develop appropriate and feasible designs.

Appendix A

Proof of Theorem 1

Let $U = min(T, C)$ and λ_i is the hazard of the ith sample, $i = 1, 2$. Since

$$\frac{1}{N}I(\theta) = \frac{1}{N}\sum_{j=1}^{k}\frac{d_{j+}r_{j1}r_{j2}e^{\theta}}{(r_{j1}+r_{j2}e^{\theta})^2} = \sum_{j=1}^{k}\frac{(d_{j+}/N)(r_{j1}/N)(r_{j2}/N)e^{\theta}}{\{(r_{j1}/N)+(r_{j2}/N)e^{\theta}\}^2}.$$

in which, for $i = 1, 2$, as $n \to \infty$,

$$\frac{d_{j+}}{N} \xrightarrow{p} P(Z = i, U \geq t_{(j)}) \equiv u(t_{(j)}),\ \frac{r_{ji}}{N} \xrightarrow{p} P(U \geq t_{(j)}, Z = i-1),$$

where $\lambda_i(t) = P(T = t|Z = i-1)/P(T \geq t|Z = i-1)$. Thus with $e^{\theta} = \lambda_2(t)/\lambda_1(t), \forall t$,

$$\begin{aligned}\frac{(r_{j1}/N)(r_{j2}/N)e^{\theta}}{\{(r_{j1}/N)+(r_{j2}/N)e^{\theta}\}^2} &= \left\{\frac{(r_{j1}/N)}{(r_{j1}/N)+(r_{j2}/N)e^{\theta}}\right\}\left\{\frac{(r_{j2}/N)e^{\theta}}{(r_{j1}/N)+(r_{j2}/N)e^{\theta}}\right\}\\ &\overset{p}{\approx} \left\{\frac{P(U \geq t_{(j)}, Z=0)\lambda_1}{P(U \geq t_{(j)}, Z=0)\lambda_1 + P(U \geq t_{(j)}, Z=1)\lambda_2}\right\}\cdot\\ &\quad\left\{\frac{P(U \geq t_{(j)}, Z=1)\lambda_2}{P(U \geq t_{(j)}, Z=0)\lambda_1 + P(U \geq t_{(j)}, Z=1)\lambda_2}\right\}\end{aligned}$$

For $P(U \geq t_{(j)}, Z = i) = P(C \geq t_{(j)}|Z = i)P(T \geq t_{(j)}|Z = i)$,

$$\begin{aligned}&P(U \geq t_{(j)}, Z = i-1)\lambda_i\\ =\ &P(C > t_{(j)}|Z = i-1)P(T = t_{(j)}|Z = i-1)P(Z = i-1)\\ =\ &P(\Delta = 1, T = t_{(j)}, Z = i-1),\end{aligned}$$

and

$$\begin{aligned}&\frac{(d_{j+}/N)(r_{j1}/N)(r_{j2}/N)e^{\theta}}{\{(r_{j1}/N)+(r_{j2}/N)e^{\theta}\}^2}\\ \overset{p}{\approx}\ &P(Z = 1|\Delta = 1, T = t_{(j)})P(Z = 0|\Delta = 1, T = t_{(j)})\\ \equiv\ &p(t_{(j)})q(t_{(j)}).\end{aligned}$$

Then the asymptotic properties can be obtained following a standard approach for failure time data by using the Martingale Theory and found that

$$\frac{1}{N}I(\theta) \xrightarrow{p} \int_0^\infty p(s)q(s)u(s)ds \equiv \frac{1}{v^2(\theta)}.$$

Appendix B

Proof of Theorem 2

At $\theta = \theta^0, |\theta^0| < \theta^*$, the log-rank test (3) has asymptotic power

$$\begin{aligned}
\beta_{LR}(\theta^0) &= P(\sqrt{N}|\hat{\theta}|/v(\hat{\theta}) < \tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\hat{\theta})\right\}) \\
&\approx P(\sqrt{N}|\hat{\theta}|/v(\theta^0) < \tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right\}) \\
&= P(-\tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right\} < \sqrt{N}\hat{\theta}/v(\theta^0) < \tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right\}) \\
&\approx \Phi\left(\tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right) - \sqrt{N}\theta^0/v(\theta^0)\right\} - \\
&\quad \Phi\left(-\tilde{C}_\alpha\left\{\sqrt{N}\theta^*/v(\theta^0)\right\} - \sqrt{N}\theta^0/v(\theta^0)\right),
\end{aligned}$$

as n goes to infinity. The asymptotic power of TOST can be easily derived in similar way.

References

Chan, I. F. (1998) Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine* **17**, 1403–1413.

Cheng, S. H., Jian, J. J-M, Tsai, S. Y.C., Yen, K. L., Chu, N-M, Chan, K-Y, Tan, T-D, Cheng, J.C., Leu, S-Y, Hsieh, C-Y and Huang, A. T. (2000) Long-term survival of nasopharyngeal carcinoma following concomitant radiotherapy and chemotherapy. *Int. J. Radiation Oncology Biol. Phys.* **48**, 1323–1330.

Farrington, C. P. and Manning, G. (1990) Test statistics and sample size formulae for comparative bunomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **9**, 1447–1454.

Hsueh, H. M., Liu, J. P. and Chen, J. J. (2001) Unconditional exact tests for equivalence or non-inferiority for paired binary data. *Biometrics* **57**, 478–483.

Kang, S-H and Chen, J. J. (2000) An approximate unconditional test of non-inferiority between two proportions. *Statistics in Medicine* **19**, 2089–2100.

Wellek, S. (1993) A log-rank test for equivalence of two survivor functions. *Biometrics* **49**, 877–881.

Table 1-1. With $\delta = .15$, empirical power $\hat{\beta}$ / asymptotic power β, of Wellek's test, TOST and the noninferiority test in a log-normal case : $S_1(t) = \Phi(2 - \ln(t))$.

				Size				Power	
				$\theta^0 = .4106$		$\theta^0 = -.4106$		$\theta^0 = 0$	
N	n_2	n_1		$\hat{\beta}$	β	$\hat{\beta}$	β	$\hat{\beta}$	β
50	25	25	β_{LR}	.0509	.0501	.0481	.0501	.1146	.1168
			β_{TOST}	.0000	.0000	.0000	.0000	.0000	.0000
			β_{NI}	.0528	.0500	-	-	.3566	.3678
75	50	25	β_{LR}	.0529	.0501	.0457	.0500	.1465	.1542
			β_{TOST}	.0000	.0000	.0000	.0000	.0000	.0000
			β_{NI}	.0576	.0500	-	-	.4459	.4461
100	50	50	β_{LR}	.0486	.0500	.0505	.0500	.2445	.2597
			β_{TOST}	.0303	.0305	.0230	.0225	.1344	.1614
			β_{NI}	.0505	.0500	-	-	.5660	.5807
125	75	50	β_{LR}	.0511	.0500	.0476	.0500	.3241	.3420
			β_{TOST}	.0447	.0419	.0355	.0387	.2690	.2961
			β_{NI}	.0526	.0500	-	-	.6395	.6481
150	75	75	β_{LR}	.0509	.0500	.0506	.0500	.4558	.4758
			β_{TOST}	.0489	.0481	.0472	.0467	.4414	.4641
			β_{NI}	.0505	.0500	-	-	.7204	.7321
175	100	75	β_{LR}	.0549	.0500	.0507	.0500	.5453	.5658
			β_{TOST}	.0543	.0493	.0498	.0488	.5417	.5619
			β_{NI}	.0553	.0500	-	-	.7649	.7809
200	100	100	β_{LR}	.0485	.0500	.0494	.0500	.6485	.6684
			β_{TOST}	.0484	.0498	.0489	.0497	.6475	.6676
			β_{NI}	.0488	.0500	-	-	.8217	.8338
250	125	125	β_{LR}	.0523	.0500	.0516	.0500	.7847	.7988
			β_{TOST}	.0523	.0500	.0516	.0500	.7847	.7987
			β_{NI}	.0523	.0500	-	-	.8940	.8994
300	150	150	β_{LR}	.0516	.0500	.0556	.0500	.8707	.8805
			β_{TOST}	.0516	.0500	.0556	.0500	.8707	.8805
			β_{NI}	.0516	.0500	-	-	.9325	.9402

Table 1-2. With $\delta = .1$, empirical power $\hat{\beta}$ / asymptotic power β, of Wellek's test, TOST and the noninferiority test in a log-normal case : $S_1(t) = \Phi(2 - \ln(t))$.

				Size				Power	
				$\theta^0 = .2727$		$\theta^0 = -.2727$		$\theta^0 = 0$	
N	n_2	n_1		$\hat{\beta}$	β	$\hat{\beta}$	β	$\hat{\beta}$	β
100	50	50	β_{LR}	.0491	.0500	.0466	.0501	.0993	.1058
			β_{TOST}	.0000	.0000	.0000	.0000	.0000	.0000
			β_{NI}	.0503	.0500	-	-	.3323	.3382
200	100	100	β_{LR}	.0473	.0500	.0522	.0500	.2070	.2173
			β_{TOST}	.0156	.0177	.0104	.0103	.0591	.0726
			β_{NI}	.0450	.0500	-	-	.5291	.5363
300	150	150	β_{LR}	.0503	.0500	.0516	.0500	.3839	.3967
			β_{TOST}	.0469	.0458	.0454	.0441	.3543	.3697
			β_{NI}	.0530	.0500	-	-	.6764	.6848
400	200	200	β_{LR}	.0491	.0500	.0491	.0500	.5800	.5852
			β_{TOST}	.0483	.0495	.0484	.0492	.5763	.5822
			β_{NI}	.0490	.0500	-	-	.7918	.7911
500	250	250	β_{LR}	.0497	.0500	.0518	.0500	.7268	.7289
			β_{TOST}	.0496	.0499	.0518	.0499	.7266	.7287
			β_{NI}	.0496	.0500	-	-	.8629	.8643
600	300	300	β_{LR}	.0490	.0500	.0506	.0500	.8254	.8268
			β_{TOST}	.0490	.0500	.0506	.0500	.8254	.8268
			β_{NI}	.0490	.0500	-	-	.9111	.9134
700	350	350	β_{LR}	.0502	.0500	.0511	.0500	.8839	.8910
			β_{TOST}	.0502	.0500	.0511	.0500	.8839	.8910
			β_{NI}	.0502	.0500	-	-	.9425	.9455

Table 2-1 With $\delta = .15$, empirical power $\hat{\beta}$ / asymptotic power β, of Wellek's test, TOST and the noninferiority test in an exponential model: $S_1(t) = exp(-t/\lambda_1)$, where λ_1 is selected such that $S_1(5) = 0.55$, and the censoring distribution $S_{C1}(t) = S_{C2} = exp(-t/\lambda_c)$, where λ_c is selected such that the censoring rate is 20%.

				Size				Power	
				$\theta^0 = .4106$		$\theta^0 = -.4106$		$\theta^0 = 0$	
N	n_2	n_1		$\hat{\beta}$	β	$\hat{\beta}$	β	$\hat{\beta}$	β
50	25	25	β_{LR}	.0461	.0501	.0520	.0501	.1063	.1155
			β_{TOST}	.0000	.0000	.0000	.0000	.0000	.0000
			β_{NI}	.0526	.0500	-	-	.3485	.3645
100	50	50	β_{LR}	.0529	.0500	.0495	.0500	.2401	.2548
			β_{TOST}	.0315	.0293	.0200	.0210	.1317	.1518
			β_{NI}	.0515	.0500	-	-	.5735	.5759
150	75	75	β_{LR}	.0482	.0500	.0525	.0500	.4504	.4671
			β_{TOST}	.0460	.0479	.0494	.0464	.4347	.4542
			β_{NI}	.0479	.0500	-	-	.7140	.7762
200	100	100	β_{LR}	.0539	.0500	.0489	.0500	.6399	.6599
			β_{TOST}	.0529	.0498	.0484	.0496	.6378	.6589
			β_{NI}	.0539	.0500	-	-	.8200	.8295
250	125	125	β_{LR}	.0469	.0500	.0523	.0500	.7722	.7919
			β_{TOST}	.0469	.0500	.0523	.0500	.7722	.7918
			β_{NI}	.0469	.0500	-	-	.8871	.8959
300	150	150	β_{LR}	.0480	.0500	.0510	.0500	.8622	.8754
			β_{TOST}	.0480	.0500	.0509	.0500	.8622	.8754
			β_{NI}	.0480	.0500	-	-	.9331	.9377
350	175	175	β_{LR}	.0492	.0500	.0517	.0500	.9219	.9266
			β_{TOST}	.0492	.0500	.0517	.0500	.9219	.9266
			β_{NI}	.0492	.0500	-	-	.9616	.9633

Table 2-2 With $\delta = .1$, empirical power $\hat{\beta}$ / asymptotic power β of Wellek's test, TOST and the noninferiority test in a exponential model: $S_1(t) = exp(-t/\lambda_1)$, where λ_1 is selected such that $S_1(5) = 0.55$, and the censoring distribution $S_{C1}(t) = S_{C2} = exp(-t/\lambda_c)$, where λ_c is selected such that the censoring rate is 20%.

				Size				Power	
				$\theta^0 = .2727$		$\theta^0 = -.2727$		$\theta^0 = 0$	
N	n_2	n_1		$\hat{\beta}$	β	$\hat{\beta}$	β	$\hat{\beta}$	β
100	50	50	β_{LR}	.0482	.0501	.0479	.0501	.1019	.1049
			β_{TOST}	.0000	.0000	.0000	.0000	.0000	.0000
			β_{NI}	.0506	.0500	-	-	.3251	.3353
200	100	100	β_{LR}	.0526	.0500	.0456	.0500	.2094	.2136
			β_{TOST}	.0182	.0160	.0071	.0084	.0510	.0635
			β_{NI}	.0539	.0500	-	-	.5315	.5317
300	150	150	β_{LR}	.0501	.0500	.0473	.0500	.3719	.3888
			β_{TOST}	.0447	.0454	.0417	.0436	.3408	.3597
			β_{NI}	.0505	.0500	-	-	.6757	.6798
400	200	200	β_{LR}	.0500	.0500	.0470	.0500	.5609	.5762
			β_{TOST}	.0496	.0494	.0462	.0491	.5574	.5728
			β_{NI}	.0498	.0500	-	-	.7803	.7864
500	250	250	β_{LR}	.0512	.0500	.0482	.0500	.7223	.7210
			β_{TOST}	.0512	.0499	.0481	.0499	.7214	.7207
			β_{NI}	.0512	.0500	-	-	.8564	.8603
600	300	300	β_{LR}	.0510	.0500	.0516	.0500	.8121	.8205
			β_{TOST}	.0510	.0500	.0516	.0500	.8121	.8204
			β_{NI}	.0510	.0500	-	-	.9077	.9102
700	350	350	β_{LR}	.0475	.0500	.0485	.0500	.8821	.8862
			β_{TOST}	.0475	.0500	.0485	.0500	.8821	.8862
			β_{NI}	.0475	.0500	-	-	.9426	.9431

Table 3. Sample size requried per arm under each model with $\rho = 1/2$,

$\delta = ..15$ and $S_1(5) = .55$, censoring rate =.20

		Log-normal case			Exponential case		
β^*		$(\infty, *)^a$	$(5,1)^b$	$(5,2)^c$	$(\infty, *)^a$	$(5,1)^b$	$(5,2)^c$
..7	LRT[1]	107	302	233	107	271	224
	TOST[2]	107	302	234	107	271	224
	NIT[3]	70	198	153	70	178	147
..8	LRT	127	360	278	127	323	266
	TOST	127	360	278	127	323	266
	NIT	92	260	201	92	234	192
..9	LRT	161	454	351	161	408	336
	TOST	161	454	351	161	408	336
	NIT	127	360	272	127	323	266

$(\infty, *)^a$: infinity accrual and follow-up;

$(5,1)^b$: 5 years of uniform accrual and 1 additional year of follow up;

$(5,2)^c$: 5 years of uniform accrual and 2 additional years of follow up;

LRT[1] : Wellek's Log-Rank test;

TOST[2] : two one-sided test;

NIT[3] : non-inferiority one-sided test.