

## **SOME FUNDAMENTAL ISSUES WITH NON-INFERIORITY TESTING IN ACTIVE CONTROLLED TRIALS\***

H.M. James Hung, Sue-Jane Wang, Yi Tsong, John Lawrence, Robert T. O'Neill  
H.M. James Hung, DB1/OB/CDER/FDA, HFD-710, 15B45, 5600 Fishers Lane, Rockville, MD 20857, USA

**KEY WORDS:** efficacy, non-inferiority margin, preservation of active control effect, historical trial, alpha error

### **1. INTRODUCTION**

Non-inferiority testing for therapeutic effectiveness in active controlled clinical trials is often controversial despite the fact that there is considerable practical experience in using this approach to establish the effectiveness of an experimental therapy. The controversy surrounds many issues including the choice of active control, the selection of historical trials to estimate the effect of the active control, statistical methods for estimation of the control effect, methods for determination of the non-inferiority margin, the analysis population, and etc. [1-16].

Many other fundamental issues have not been addressed in practical applications of non-inferiority testing. For instance, what is the main objective of non-inferiority testing when designing the active control trial without a placebo arm? Is pre-specification of a non-inferiority margin always required? If so, what needs to be considered in defining it? An issue that has not been discussed for statistical inference in non-inferiority testing concerns the control of the relevant alpha error (or type I error) probability of making a false conclusion, although the evaluation of bias in the estimate of the new treatment effect is also a concern. The alpha error will be a main subject of this paper and discussed in Section 4. We will discuss the potential problems in defining the non-inferiority margin based on the estimate of the active control effect in Sections 3, 4 and 5.

### **2. OBJECTIVE OF NON-INFERIORITY TESTING**

Two major inter-related objectives regarding effectiveness of an experimental therapy are often considered in non-inferiority trials. One objective pertains to the establishment of efficacy of the experimental treatment. As articulated in Fisher et al [17] and Hasselblad and Kong [18], through a direct randomized comparison of the experimental treatment with the active control treatment, it is inferred from non-inferiority testing that the experimental treatment would have been more effective than placebo had a placebo been included in the trial. The other objective is

to infer that the experimental treatment is not much less effective than the active control (notion of “not much inferior”) by a specified margin. Achieving this latter objective will be meaningful only if the former objective is achieved because if the margin is too wide, an ineffective therapy can be erroneously judged effective. As pointed out by Siegel [13] and Fleming [14], an experimental treatment similar to or somewhat less effective than the active control may not be desirable unless it has other significant benefits (e.g., less toxic, less costly, easier to administer, etc.) that outweigh the seeming loss of efficacy. Hence, defining quantitatively the concept of “not much inferior” is quite a difficult task. A single threshold value (called non-inferiority margin) needs to be determined in order to define this concept and enable one to infer the efficacy of the experimental treatment provided that the safety profile and other factors are well justified. The magnitude, the variability and constancy of the effect of the active control are some of the major factors for consideration in determination of the non-inferiority margin.

Recently, an approach used in non-inferiority testing (e.g., CBER/FDA Memorandum [11]) requires that the experimental therapy preserve some fraction of the effect of the active control therapy. Preservation of some fraction of the control effect is important to ensure that the efficacy of the experimental therapy relative to placebo can be established with great confidence. In some cases, it may be sufficient to ensure that the amount of loss of therapeutic effect with the experimental treatment is that maximally allowable. However, the concept of “not much inferior” cannot always be defined by a quantitative retention of some fraction of the control effect.

### **3. NON-INFERIORITY MARGIN**

The traditional framework for non-inferiority testing (e.g., Blackwelder [19]) requires that the value of the non-inferiority margin be set in the hypothesis. To illustrate the idea, let  $T$ ,  $C$ , and  $P$  denote the incidence rates of a clinical event associated with the experimental treatment, the control treatment, and the placebo, respectively, in the patient population targeted by the active control study. Let  $C_0$  and  $P_0$  be the event probabilities for the control treatment and placebo, respectively, in the historical trial populations. Note that  $C$  and  $C_0$  may differ and so may  $P$  and  $P_0$  because of potential heterogeneity in the trial populations, possible changes in event rates over time,

---

\* The reviews expressed in this article do not represent those of the U.S. Food and Drug Administration.

other sources of heterogeneity, etc. If the treatment effect is expressed in terms of relative risk, the hypotheses for non-inferiority testing can be formulated as follows:

$$H_0: T/C \geq \delta \quad \text{versus} \quad H_1: T/C < \delta,$$

where  $A/B$  is the relative risk of  $A$  versus  $B$  and  $\delta$  is the non-inferiority margin, a parameter usually greater than one but not much greater. The quantity  $(\delta - 1)$  is the amount of excess risk allowable for the experimental treatment over the control. To define the concept of “not much inferior”, the margin  $\delta$  needs to reflect the maximal allowable loss of efficacy associated with the experimental treatment relative to the active control given the consideration of its benefits and risks, and hence  $\delta$  can depend on many factors besides the value of the control effect.

To preserve greater than  $100\lambda\%$  or, equivalently, allow a loss of less than  $100(1-\lambda)\%$  of the active control effect by the experimental treatment in terms of relative risk, we have  $1 - T/P > \lambda(1 - C/P)$ , equivalently,  $T/C < 1 + (1-\lambda)(P/C - 1)$ , where  $\lambda$  is a pre-specified fixed constant between 0 and 1 inclusive. Alternatively, the hypotheses that the experimental treatment preserves  $>100\gamma\%$  of the control effect on the log relative risk scale are

$$\begin{aligned} K_0: \{\log(P) - \log(T)\}/\{\log(P) - \log(C)\} \leq \gamma \quad \text{vs.} \quad K_1: \{\log(P) - \log(T)\}/\{\log(P) - \log(C)\} > \gamma, \\ \text{equivalently, } K_0: \log(T) - \log(C) \geq \delta \quad \text{vs.} \\ K_1: \log(T) - \log(C) < \delta, \end{aligned} \quad (1)$$

where  $\delta = (1-\gamma)\{\log(P) - \log(C)\}$  is the non-inferiority margin. Often it is more desirable to work on the log relative risk scale when relative risk, odds ratio or hazard ratio is used to explain the risk of occurrence of the adverse clinical event. One reason is that the statistics on this scale are better approximated by the Gaussian distribution. The relationship between the preservation on the risk ratio scale and the preservation on the log relative risk scale has been explored in Wang, Hung and Tsong [20]. For  $\gamma = 0$ , the hypothesis  $K_1$  is simply that the experimental treatment is superior to the putative placebo in effectiveness. For  $\gamma = 1$ ,  $K_1$  is that the experimental treatment is more effective than the active control.

In the hypothesis of the effect retention as in (1), the non-inferiority margin  $\delta$  is an explicit function of the control effect  $\log(P) - \log(C)$  in the active control trial population. But the value of  $\delta$  is unknown and the hypothesis (1) is generally not testable. It can be tested if the active control effects are equal in the active control trial patient population and the historical trial patient populations (this is the so-called ‘Constancy Condition’) and if the historical data can provide an unbiased estimate for the common control effect. These assumptions are often unverifiable using the data.

It is worth emphasizing that any single estimate or aggregate of the estimates of the control effect from the historical trials cannot directly define the non-inferiority statistical hypotheses of concern. As a basic frequentist statistical principle, the hypothesis of non-inferiority is formulated with population parameters, not the estimates from the studies. The relevant active control effect to be preserved is the effect on the active control trial population, not on the historical trial populations. The fact that the value of  $\delta$  is unknown and at best estimable from historical trials makes it extremely difficult to set the value of  $\delta$  to a fixed number for properly defining the non-inferiority hypothesis, as required under the traditional framework. For instance, for inferring the effect retention as described in (1), the fixed parameter value for the margin  $\delta$  requires coverage of the range of all plausible values of the active control effect which is unknown and would otherwise be directly estimable from the current active control trial if it included the placebo. This is quite a formidable task and has tremendous implication on the statistical error of making a false inference as to be discussed in the following sections.

#### 4. PROBABILITY OF FALSELY CONCLUDING PERCENT RESERVATION

##### Confidence Interval Method

When the non-inferiority margin  $\delta$  in (1) is a fixed known constant, a natural method for testing the percent effect preservation is the conventional confidence interval approach requiring that the two-sided  $100(1-2\alpha)\%$  confidence interval of  $\log(T/C)$  lies below  $\delta$ . It is straightforward to show that in this case the alpha error (or type I error) probability of falsely concluding that the experimental treatment preserves greater than  $100\gamma\%$  of the control effect with this approach is at most  $\alpha$  (conventionally set at 2.5%). However, when  $\delta$  is estimated from the historical trials, even if the constancy condition holds, the uncertainty underlying the estimate of  $\delta$  needs to be incorporated in the calculation of the alpha error probability.

Given  $\gamma$ , the estimate of  $\delta$  can be obtained from the historical trials if the constancy condition holds. For instance, if the lower limit of the 95% confidence interval of the historical trial estimate  $\log(\tilde{P}_0) - \log(\tilde{C}_0)$  is used to estimate the control effect  $\log(P) - \log(C)$ , then it follows that the confidence interval approach rejects  $K_0$  if

$$\begin{aligned} & \log(\hat{T}) - \log(\hat{C}) + 1.96 \sigma_{TC} \\ & < (1 - \gamma)(\log(\tilde{P}_0) - \log(\tilde{C}_0)) - 1.96 \sigma_{PC0} \end{aligned}$$

where  $\log(\hat{T}) - \log(\hat{C})$  is the estimate of  $\log(T) - \log(C)$  from the active control trial with its standard error  $\sigma_{TC}$  and  $\sigma_{PC0}$  is the standard error of  $\log(\tilde{P}_0) - \log(\tilde{C}_0)$ .

If the uncertainty underlying the estimate of  $\delta$  is incorporated in the assessment of statistical error, then under  $K_0$  the maximum probability for this rejection region is  $\Phi(-1.96f)$ , where  $\Phi$  is the standard normal distribution and

$f = \{\sigma_{TC} + (1-\gamma)\sigma_{PC0}\} / \{\sigma_{TC}^2 + (1-\gamma)^2\sigma_{PC0}^2\}^{1/2}$  which is always greater than one. Thus, the alpha error probability of the confidence interval approach using the lower limit of 95% confidence interval to define the margin is always conservative (less than 2.5%), when the constancy condition holds.

Alternatively, if the point estimate  $\log(\tilde{P}_0) - \log(\tilde{C}_0)$  is used to define the margin, then the rejection region for  $K_0$  based on this confidence interval approach is

$$\log(\hat{T}) - \log(\hat{C}) + 1.96\sigma_{TC} < (1-\gamma)(\log(\tilde{P}_0) - \log(\tilde{C}_0))$$

Under  $K_0$ , the maximum probability associated with this rejection region is  $\Phi(-1.96h)$ , where  $h = \sigma_{TC} / \{\sigma_{TC}^2 + (1-\gamma)^2\sigma_{PC0}^2\}^{1/2}$  which is positive and always smaller than one. Thus the alpha error probability of the confidence interval approach using the point estimate of the control effect to define the margin can be anti-conservative (greater than 2.5%), even when the constancy condition holds.

The problems of conservatism and anti-conservatism discussed above with the confidence interval approach naturally lead one to search for an estimated margin  $\delta^*$  such that the confidence interval approach using this margin has the alpha error attained exactly at 2.5% level, that is, on the boundary of  $K_0$ ,

$$\Pr(\log(\hat{T}) - \log(\hat{C}) + 1.96\sigma_{TC} < \delta^*) = 0.025.$$

By some algebraic manipulations, one can show that

$$\delta^* = -1.96\{\sqrt{\sigma_{TC}^2 + (1-\gamma)^2\sigma_{PC0}^2} - \sigma_{TC}\} + (1-\gamma)(\log(\tilde{P}_0) - \log(\tilde{C}_0)),$$

which is an explicit function of the estimated active control effect and its standard error  $\sigma_{PC0}$  from historical trials.

The estimated margin  $\delta^*$  is also a function of the standard error  $\sigma_{TC}$  of the effect of the experimental therapy relative to the control from the active control trial. That is, the estimated margin  $\delta^*$  depends on the sample size (or statistical information) of the current active control trial. As the basic principle of designing an experiment, the sample size is to be planned for testing non-inferiority with a pre-specified target margin at certain alpha error rate and power. Now the non-inferiority margin  $\delta^*$  which must be pre-specified prior to sample size planning needs to be determined

using the sample size yet to be planned. Such a selection of the non-inferiority margin is irrelevant and violates the principle of design of experiment. Moreover, as analytically proven in the Appendix,  $\delta^*$  increases as  $\sigma_{TC}$  increases. That is, the larger is the sample size of the planned active control trial, the tighter the non-inferiority margin  $\delta^*$  is required. This determination of the non-inferiority margin creates an illogical problem.

### Preservation Test Method

The formulation in (1) presents another way for incorporating the uncertainty with the estimate of  $\log(P/C)$  obtained from historical trials. If the constancy condition holds, a test method can be developed for the hypotheses such as (1). References include Holmgren [21] and Wang, Hung and Tsong [20]. A Bayesian analogue of this method was given by Simon [22]. One of the test methods is the preservation test  $Z_{pv}$  studied by Wang, Hung and Tsong and given by

$$Z_{pv} = \frac{\log(\hat{T}) - \log(\hat{C}) - (1-\gamma)(\log(\tilde{P}_0) - \log(\tilde{C}_0))}{\sqrt{\sigma_{TC}^2 + (1-\gamma)^2\sigma_{PC0}^2}}.$$

Based on the convention of 2.5% type I error probability, the rejection region of this test for  $K_0$  is  $Z_{pv} < -1.96$ , provided that the normal approximation is valid. If the constancy condition holds, the alpha error probability associated with this method is always less than or equal to 2.5% and achieves 2.5% on the boundary of  $K_0$  regardless of the true value of the event rates  $P$  and  $C$ .

### Impact of Violation of Constancy Assumption

The preservation test method is known to be highly sensitive to the constancy condition, as reported by Wang, Hung and Tsong. In particular, if the effect of the active control is worse in the active control trial population as compared to the historical trial populations, this approach is invalid in the sense that the alpha error probability exceeds the target level of 2.5%. The confidence interval approach using either the point estimate of the control effect to define the margin or the estimated margin  $\delta^*$  is either worse or no better than the preservation test method in terms of the alpha error probability. In contrast, the confidence interval approach using the worst limit of the 95% confidence interval of the active control effect to define the margin is much less sensitive to the constancy condition.

### Conditional Versus Unconditional Alpha Errors

The alpha error probability discussed above is an unconditional error in the sense that the calculation of this error incorporates the statistical uncertainty underlying the estimation for the non-inferiority margin. This unconditional error involves cross-trial statistical inference beyond that within the active control trial at stake. The traditional alpha error probability is a conditional error; that is, the margin is treated as if it

were a fixed known number after it is estimated. The calculation of the conditional error involves only the statistical uncertainty within the active control trial and does not embed the uncertainty associated with the estimation for the non-inferiority margin. The preservation test can only be performed in the context of unconditional alpha error. But the confidence interval approach can be used in the context of conditional or unconditional error.

When the non-inferiority margin is estimated, with the confidence interval approach, the conditional alpha error cannot offer a measure of confidence level that is always required in assessing statistical evidence. The reason is, as noted above, that the value of  $\delta$  is unknown and at best estimable from historical trials and thus one will never be sure that the fixed number set for the margin  $\delta$  is tighter than the intended margin that depends on the unknown value of the active control effect. One may attempt to use a very conservative non-inferiority margin with which the result of the confidence interval method hopefully leads one to conclude non-inferiority with a sense of small risk or more precisely a sense of comfort. This risk is beyond the conditional alpha error and not quantifiable statistically under the frequentist framework.

## 5. EXAMPLE

To illustrate the implication of the discussion above, consider the scenario where the result of a meta-analysis of historical trials is used to estimate the effect of an active control on a clinical event in a target patient population with a cardiovascular disease. Suppose that the estimated 5-year event rates in the active control and the placebo groups are 14% and 18%, respectively, based on a total of 5,000 patients per group. The estimated relative risk of the active control relative to the placebo is 0.78 with 95% confidence interval of (0.71, 0.85). Suppose that this active control is selected as the comparator in an active-controlled non-inferiority trial for studying a new treatment and the relative risk of the active control from the historical trial populations is assumed applicable to the active controlled trial population (that is, the constancy condition is met).

To test for 50% preservation of the control effect in terms of log relative risk, the non-inferiority margin defined using the estimated relative risk 0.78 is 0.13. Note that this margin is an estimate from historical trials. With the confidence interval approach, if the upper limit of 95% confidence interval for the log relative risk of the new treatment versus the active control is smaller than the margin, one would conclude that the new treatment preserves 50% of the control effect. However, the unconditional alpha error probability of incorrectly asserting 50% preservation is greater than 2.5%. As illustrated in Table 1, the error

probability can be more than double 2.5%. Thus, though the conditional alpha error probability is set at 2.5%, one can only conclude that the excess log relative risk of the new treatment over the active control is smaller than 0.13 but cannot conclude at 2.5% level of significance that the new treatment preserves 50% of the control effect.

In contrast, the non-inferiority margin defined using the worst limit 0.85 of the 95% confidence interval is 0.081. This margin is also an estimate but the unconditional alpha error probability of falsely concluding 50% preservation is now in fact smaller than 2.5% as proven in Section 4. It is very conservative, e.g., in the order of  $10^{-3}$ , as shown in Table 1.

The unconditional alpha error probability associated with the confidence interval approach will vary as the sample size and/or the values of the nuisance parameters in the active controlled trial change. Hence, it is not possible to exactly attain 2.5% in unconditional alpha error of the confidence interval approach regardless of the sample size of the active control trial. In contrast, the preservation test  $Z_{pv}$  always has a correct alpha error probability, that is, 2.5%, regardless of the sample size and the values of the nuisance parameters, if the constancy condition holds.

The comparisons among these approaches discussed above are made under the assumption that the constancy condition holds, which is often unverifiable and assumed away. If this assumption is violated, statistical inference for non-inferiority based on either preservation test or confidence interval method may not be interpretable because the statistical error of drawing a false assertion may be uncontrolled, as pointed out by Wang, Hung and Tsong [20].

## 6. DISCUSSION

Non-inferiority is a term, as many perceive, meant to capture the concept of “not much inferior” in terms of a quantitative level of therapeutic effectiveness. In recent practice, the main objective frequently articulated at the planning stage of a non-inferiority trial pertains to evaluating the efficacy of the experimental therapy relative to a putative placebo or to the proportion of the effect of the active control that may be preserved by the experimental therapy. Achieving the latter objective is intended to lead one to conclude that the experimental therapy is therapeutically efficacious (i.e., would have beaten placebo had placebo been compared against). These objectives should be distinguished in designing an active control trial without a placebo arm.

The pre-specification of a non-inferiority margin is definitely needed if the trial objective is to demonstrate that the effect of the experimental therapy is not much

less than that of the active control. In this case the determination of the margin would depend on many factors, such as, the magnitude of the control effect, safety profiles of the two comparative therapies, administration factor, cost factor, and etc. Thus, the margin selection cannot be based purely on some statistical criteria because the uncertainty associated with the selected margin is not statistically measurable.

If the constancy condition holds and there is no bias from the historical trial data, the pre-specification of a non-inferiority margin is arguably unnecessary if the goal of non-inferiority testing is to demonstrate the efficacy of the experimental treatment or the preservation of some percent of the active control effect. In practice it may be preferred to set the margin when testing for effect preservation. We need to be reminded of the fact that the active control effect is unknown even though it may be estimated from historical trials if the constancy condition holds. The variability surrounding this estimate needs to be considered when attaching a statistical error of falsely asserting that the experimental therapy preserves greater than 50%, say, of the active control effect. This consideration necessitates the distinction between conditional and unconditional alpha errors for making such a false assertion. The conditional error is evaluated by treating the estimated margin as if it were true (fixed and known) while the calculation of the unconditional error incorporates the measure of statistical uncertainty with the estimated margin.

For testing effect retention, the preservation test method does not set the value of the non-inferiority margin and needs only pre-specification of percent preservation. If the constancy assumption holds, the statistical alpha error probability of falsely concluding a given percent of effect retention with this approach will attain exactly the target level regardless of the value of the control effect, the value of nuisance parameter and sample size to be planned in the non-inferiority trial. In contrast, the classical confidence interval approach that needs pre-specification of a fixed non-inferiority margin cannot attain the target alpha error level since the true effect of the control is not known. For instance, if the point estimate of the control effect obtained from historical trials is used to define the margin, then the confidence interval approach will always have alpha error probability exceeding the target level. On the other hand, if the worst limit of the confidence interval of the control effect is used, then the alpha error is always smaller than the target level. Furthermore, if the margin is estimated such that the alpha error meets the target level, then we have exposed the problem of such a margin. Such a margin selection is irrelevant.

When the constancy assumption is very much in doubt, both the conditional and the unconditional alpha

errors may be irrelevant. In this case, one will have to avoid the non-inferiority testing unless one is able to select an extremely conservative margin so that the result of the preservation test is interpretable with a very small risk of drawing a false conclusion. Such a risk is beyond the traditional alpha error and not quantifiable statistically.

#### ACKNOWLEDGEMENTS

This research was supported by RSR fund #01-20 of Center for Drug Evaluation and Research, Food and Drug Administration. Thanks are also due to Dr. Lu Cui for sharing his views before he left FDA and to Dr. Charles Anello for his comments.

#### REFERENCES

1. Temple, R. (1983). 'Difficulties in evaluating positive control trials', *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 1-7.
2. Fleming, T.R. (1987). 'Treatment evaluation in active control studies', *Cancer Treatment Reports*, 71, 1061-1064.
3. Pledger, G., Hall, D.B. (1990). 'Active control equivalence studies: do they address the efficacy issue?', *Statistical Issues in Drug Research and Development*, Marcel Dekker, New York, pp. 226-238.
4. Jones, B., Jarvis, P., Lewis, J.A., Ebbutt, A.F. (1996). 'Trials to assess equivalence: the importance of rigorous methods', *British Medical Journal*, 313, 36-39.
5. Temple, R. (1996). 'Problems in interpreting active control equivalence trials', *Accountability in Research*, 4, 267-275.
6. Wang, S.J., Hung, H.M.J., Tsong, Y., Cui, L., Nuri, W. A. (1997). 'Changing the study Objective in clinical trials', *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 64-69.
7. Ebbutt, A.F., Frith, L. (1998). 'Practical issues in equivalence trials', *Statistics in Medicine*, 17, 1691-1701.
8. Rohmel, J. (1998). 'Therapeutic equivalence investigations: statistical considerations', *Statistics in Medicine*, 17, 1703-1714.
9. Internal Conference on Harmonisation: statistical principles for clinical trials (ICH E-9), Food and Drug Administration, DHHS, 1998.
10. Internal Conference on Harmonisation: guidance on choice of control group in clinical trials (ICH E-10), Food and Drug Administration, DHHS, 1999.
11. CBER/FDA Memorandum. Summary of CBER considerations on selected aspects of active controlled trial design and analysis for the evaluation of thrombolytics in acute MI, June 1999.

12. Committee for Proprietary Medicinal Products (CPMP). 'Points to consider on switching between superiority and non-inferiority', 2000. <http://www.eudra.org/emea.html>.

13. Siegel, J.P. (2000). 'Equivalence and noninferiority trials', *American Heart Journal*, 139, S166-S170.

14. Fleming, T.R. (2000). 'Design and interpretation of equivalence trials', *American Heart Journal*, 139, S171-S176.

15. 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, Part 2: ethical and scientific issues', *Annals of Internal Medicine*, 133, 455-463.

16. 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.

17. Fisher, L.D., Gent, M., Büller, H.R. (2001). 'Active-control trials: How would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo', *American Heart Journal*, 141, 26-32.

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

19. Blackwelder, W.C. (1982). 'Proving the Null Hypothesis in clinical trials', *Controlled Clinical Trials*, 3, 345-353.

20. Wang, S.J., Hung, H.M.J., and Tsong, Y. (2001). 'Utility and pitfall of some statistical methods in active controlled clinical trials', *Controlled Clinical Trials*, 22:1-14.

21. 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.

22. Simon, R. (1999). 'Bayesian design and analysis of active control clinical trials', *Biometrics*, 55, 484-487.

### Appendix

Let  $\delta^*$  be the estimated non-inferiority margin such that on the boundary of  $K_0$ ,

$$\Pr\{\log(\hat{T}) - \log(\hat{C}) + 1.96\sigma_{TC} < \delta^*\} = 0.025.$$

Then we have

$$\delta^* = -1.96\left\{\sqrt{\sigma_{TC}^2 + (1-\gamma)^2\sigma_{PC0}^2} - \sigma_{TC}\right\} + (1-\gamma)(\log(\tilde{P}_0) - \log(\tilde{C}_0)).$$

Since the first derivative of  $\delta^*$  with respect to  $\sigma_{TC}$  is

$$-1.96\left[\frac{\sigma_{TC}}{\sqrt{\sigma_{TC}^2 + (1-\gamma)^2\sigma_{PC0}^2}} - 1\right] > 0,$$

$\delta^*$  is an increasing function of  $\sigma_{TC}$ . Moreover, by L'Hôpital's rule of Calculus, as  $\sigma_{TC} \rightarrow \infty$ ,

That is, as  $\sigma_{TC} \rightarrow \infty$ ,  $\delta^*$  will be the non-inferiority margin defined using the point estimate of the control effect of the historical control trials. In contrast, if  $\sigma_{TC} \rightarrow 0$ ,  $\delta^*$  will be the non-inferiority margin defined using the 95% confidence interval lower limit of the control effect of the historical control trials.

**Table 1. Unconditional alpha error probability of falsely concluding that the new treatment preserves 50% of the control effect (based on 100,000 replications)**

True event rate		Sample size for active control trial	Inferential method	Alpha error probability <sup>a</sup>
Placebo	Control			
18%	14%	10000/arm	CI method using point estimate	0.058
			CI method using LL <sup>b</sup> of 95% CI	0.003
			Z <sub>pv</sub> test	0.025
18%	14%	7500/arm	CI method using point estimate	0.050
			CI method using LL <sup>b</sup> of 95% CI	0.003
			Z <sub>pv</sub> test	0.025
19%	15%	10000/arm	CI method using point estimate	0.056
			CI method using LL <sup>b</sup> of 95% CI	0.003
			Z <sub>pv</sub> test	0.025

<sup>a</sup> target level is 0.025; it is liberal if > 0.025, valid if = 0.025, conservative if < 0.025

<sup>b</sup> LL: lower limit of CI (confidence interval)