Non-Constancy Adjustment for Non-Inferiority Testing in Active-Controlled Clinical Trials

Carl DiCasoli[1], Cynthia DeSouza[1], Emily Martin[1]
[1]Vertex Pharmaceuticals, 130 Waverly Street, Cambridge, MA 02139

Abstract

In new studies of oral regimens for the treatment of Hepatitis C, a potential challenge is to derive a valid method for testing non-inferiority. One critical assumption in non-inferiority trials is constancy; that is, the effect of the active control in the historical trial population is similar to the effect in the active control trial population. This assumption is at risk due to the potential heterogeneity between trial populations primarily related to subject characteristics, and secondarily to other sources of heterogeneity resulting from differences in patient management (e.g., usage of concomitant medications). To investigate the impact of the constancy assumption, we propose an adaptive two-stage method for non-inferiority testing based on a constancy adjustment followed by sample size re-estimation. We will evaluate the overall magnitude of the alpha and beta errors when implementing this two-step approach for non-inferiority testing, compared to the standard synthesis and confidence-interval approaches.

Key Words: Covariate adjustment, non-inferiority, constancy, risk difference, adaptive, interim analysis, sample size re-estimation, group-sequential, synthesis, fixed margin, alpha error, beta error, TACT method.

1. Introduction

In a recent Phase III clinical trial, the primary objective was to establish non-inferiority of a test treatment, *T*, versus an active control treatment, *C,* with respect to sustained viral response (SVR), where high values of SVR are desirable. In addition, it is expected that the test treatment is more effective than placebo treatment, *P*. However, due to ethical concerns, the placebo is not used in the active control, non-inferiority trial. Instead, historical data from a similar, previous trial of the active control versus placebo (historical control, C_0 versus historical placebo, P_0) is used to demonstrate efficacy of the test treatment relative to placebo via cross-trial inference. It is assumed that the response rate of the putative placebo in the active control trial equals the historical placebo; that is, $P=P_0$. Furthermore, the non-inferiority trial assumes constancy of the active control effect as in the historical trial (i.e., $C-P \approx C_0-P_0$). If this constancy assumption is violated, new methodology will be required to salvage the active control trial.

The current methodology for non-inferiority trials utilizes two well-known methods: the synthesis method and the conservative confidence interval (or fixed margin) method.

At the design stage, the conservative confidence interval takes into account information regarding M1, which is the lower bound of the historical active control effect and M2 is the portion of the efficacy of the active control that is not preserved in the efficacy of the test treatment; it is called the non-inferiority margin and is denoted by $\boldsymbol{\delta_{initial}}$. Specifically,

$$\mathbf{M2 = (1-\eta)\ M1,} \qquad \mathbf{(1)}$$

where $\mathbf{M1= \{C_0 - P_0 - z_{0.025}\,\sigma_{P0C0}\}}$, $\boldsymbol{\eta}$ represents a preservation level ranging between 0 and 1, and $\boldsymbol{\sigma_{P0C0}}$ represents the standard deviation of $C_0 - P_0$ from the historical trial

At the analysis stage, the test for non-inferiority is based on showing that the upper bound of the 95% confidence interval for the difference between C and T is within the specified non-inferiority margin, $\boldsymbol{\delta_{initial}}$. That is,

$$\boldsymbol{C - T + z_{0.025}\,\sigma_{TC} < (1-\eta)\{C_0 - P_0 - z_{0.025}\,\sigma_{P0C0}\},} \qquad \mathbf{(2)}$$

where $\boldsymbol{\sigma_{TC}}$ represents the standard deviation of $C - T$ in the non-inferiority trial. Overall, the fixed margin method is conservative in controlling the type I error but may not be efficient in terms of controlling the Type II error.

In contrast, the synthesis method, at the analysis stage, "synthesizes" or combines the test treatment effect relative to the active control along with the estimate of the active control effect from the historical trial in such a way that it can be used to test non-inferiority. The synthesis method treats both sources of data as if they are from the same randomized trial, omitting trial-to-trial variability. This could potentially lead to underestimating the standard error and result in a higher chance of committing a Type I error. From the synthesis method, a single confidence interval is obtained for testing that the test treatment preserves a fixed portion of the active control effect. If the constancy assumption is violated, using the synthesis method, as compared to the fixed margin method, could result in a Type I error inflation but also greater efficiency; that is, a lower Type II error. A test statistic for the synthesis method is expressed as

$$\boldsymbol{Z_{pv}=[C - T - (1-\lambda)\{C_0 - P_0\}] \,/\, [\mathrm{sqrt}(\sigma^2_{TC} + (1-\lambda)^2\,\sigma^2_{P0C0})],} \quad \mathbf{(3)}$$

where $\boldsymbol{\lambda}$ represents a preservation level usually taken as 0.5, with range between 0 and 1, $\boldsymbol{\sigma^2_{TC}}$ represents the variance of $C - T$ in the non-inferiority trial and $\boldsymbol{\sigma^2_{P0C0}}$ represents the variance of $C_0 - P_0$ in the historical trial.

2. The TACT (Two-Stage Active Control Testing) Method

The TACT method of Wang and Hung (2003) is a two-stage group sequential hypothesis test of constancy intended to stop the study early based on futility if constancy is rejected at the interim analysis. The synthesis and confidence interval tests may have a high Type 1 error rate if the active control effect in the non-inferiority trial is very small relative to its effect in the historical trial. The steps for the TACT method are described as follows:

1. The actual data from the historical control trial are used, with a fixed historical placebo response rate P_0 and fixed historical control response rate C_0. Proceed to the next stage only if the active control is shown to be better than placebo from the available collection of historical trials.

2. The required sample size to test non-inferiority when $T=C$ is based on the synthesis test. The formula in Wang and Hung (2003) is given by

$$N = (2(1-C) / C) / [(1-\lambda)^2[((log(C) - log(P_0)) / (z_\alpha+z_\beta))^2 - \sigma^2_{P0C0}]] \qquad (4)$$

In this formula, λ is the level of preservation and σ_{P0C0} is the standard error of $\log(P_0) - \log(C_0)$.

3. The following logistic regression model is fit:

$$Logit(p) = \mu + \beta X_{C0} + \gamma X_C + \xi X_T, \qquad (5)$$

where the response rate, p derived from this model is estimated via the logit transformation; e.g., $C_0 = \exp(\mu+\beta) / (1 + \exp(\mu+\beta))$, etc, $\mu = \text{logit}(P_0) = \text{logit}(P)$, the common placebo effect, $\beta = \text{logit}(C_0 / P)$, $\gamma = \text{logit}(C / P)$, which explains the effect of the control as compared to the placebo in the historical trial populations and active control populations, respectively. Furthermore, $\xi = \text{logit}(T / P)$, which explains the comparative effect of test treatment relative to placebo, and X_h represents an indicator variable associated with treatments T, C, and C_0. Variances and standard errors are produced using the delta method.

4. Conduct the non-inferiority trial and test for constancy via the following test statistic.

$$Z_t^* = (C_t - C_0) / \sigma^*_{CtC0}, \qquad (6)$$

where σ^*_{CtC0} represents the standard error of $C_t - C_0$, and t represents the information fraction between 0 and 1 (e.g., $t = 0.5$ corresponds to the half-way point in the study when 50% of the responses are obtained). Compare Z_t^* to the lower futility boundary L via the Lan-Demets (1983) alpha-spending function. If the lower boundary is crossed at the interim, that is, $Z_t^* < L$, the trial is stopped. If not, then repeat the test at the final analysis.

5. If all tests above are satisfied, then a decision between the fixed margin or synthesis methods for testing non-inferiority at the final analysis, will be implemented by comparing Z_1^* to the upper futility boundary U via the Lan-Demets (1983) alpha-spending function. If $Z_1^* > U$, the synthesis margin will be implemented. If $L < Z_1^* < U$, the fixed margin method will be implemented as it is more conservative with respect to Type 1 error control.

However, there are some limitations to the TACT method. Notably, there is no salvage strategy to correct for constancy at the first stage interim analysis. Furthermore, the non-inferiority test is based on the $T=C$ assumption at the design stage of the study which may be a strong assumption.

3. The Covariate-Adjustment Method

The objective of the covariate-adjustment method of Nie and Soon (2010) is to address non-constancy ($C_0 \neq C$) arising from heterogeneity between patient populations in the two trials Assuming $P=P_0$, $P - C$ from the active control is compared against $P_0 - C_0$ from the historical trial. The following model is fit on the $g^{-1}(\mu_i)$ scale **where g(.) is the link function**:

$$E(y_i) = g^{-1}(\mu_i),\ \mu_i = \alpha + \beta Z_i + \sum_{k=1}^{K}(\beta_k x_{ik} + \gamma_k x_{ik} Z_i), \qquad (7)$$

$$Var(y_i) = V\{g^{-1}(\mu_i)\},$$

and where i represents the ith subject, $Z_i = 1$ represents placebo (P), $Z_i = 0$ represents active control (C), x_{ik} is the kth covariate, $y_i = 1$ represents response, and $y_i = 0$ represents no response. β_k is the kth covariate effect and γ_k is the interaction effect of covariate x_{ik} with treatment Z_i. Notice that the treatment effect will change with the covariates x_{ik}.

If the constancy assumption is rejected, the NI margin $\boldsymbol{\delta_{adjusted}}$ is recalibrated to the active control population via the regression model in (7) and is defined as the lower bound of a (1-α)100 % CI of $P–C$, where

$$P-C = \boldsymbol{\beta} + \sum_{k=1}^{K} \boldsymbol{\gamma_k}\, \bar{\boldsymbol{x}}_{.k}, \qquad \textbf{(8)}$$

and $\bar{\boldsymbol{x}}_{.k}$ represents the mean of the active control population. The recalibrated estimate, $\boldsymbol{\delta_{adjusted}}$ is used to redefine the non-inferiority margin if the constancy assumption is violated and quantifies the impact of population difference between the historical and active control trials based on the regression equation (7). This covariate adjustment can be implemented for both the fixed margin and the synthesis approaches.

For the covariate-adjustment with fixed margin inference, T is non-inferior to C if the upper bound of the (1-α)100 % CI of C-T is smaller than $\boldsymbol{\delta_{adjusted}}$, the updated margin on the transformed scale of choice. For the covariate-adjustment with synthesis method inference, T is non-inferior to C if the upper bound of the (1-α)100 % CI of $(C\text{-}T)$-$(1\text{-}\lambda)(P\text{-}C) < 0$ on the transformed scale of choice.

4. Proposed Adaptive Two-Stage Method

An adaptive two-stage method is proposed to test for non-inferiority based on a constancy adjustment and sample size re-estimation when $T \neq C$, using the TACT and covariate-adjustment methods.

During the initial design stage of the trial, using equation (4), we will estimate the initial sample size, $\boldsymbol{SS_{initial}}$, based on the assumption that $\boldsymbol{T=C}$ and a specified $\boldsymbol{\delta_{initial}}$, the portion of effectiveness of active control that may not be preserved in the performance of the test treatment. Note that λ and $\boldsymbol{\delta_{initial}}$ have the following relationship: $\boldsymbol{\delta_{initial} = -(1-\lambda)(C / P_0)}$.

At the interim analysis, constancy will be tested using the group-sequential procedure of Lan-Demets (1983). If $\boldsymbol{Z_t^*} < L$, where $\boldsymbol{Z_t^*}$ is calculated via equation (6) based on the data at the interim analysis, significant non-constancy is present. In this case, the covariate-adjustment method will be implemented using equations (7) and (8) to arrive at the adjusted non-inferiority margin, $\boldsymbol{\delta_{adjusted}}$. The re-estimated sample size at the interim analysis will be calculated via equation (4) using the observed SVR rate from $\boldsymbol{T}$ and $\boldsymbol{C}$ that will be denoted as $\boldsymbol{T_{obs}}$ and $\boldsymbol{C_{obs}}$, respectively, in addition to the updated margin, $\boldsymbol{\delta_{adjusted}}$. On the contrary, if $\boldsymbol{Z_t^*} > L$, the re-estimated sample size will be calculated via equation (4) using the observed SVR rates from the test treatment and active control, $\boldsymbol{T_{obs}}$ and $\boldsymbol{C_{obs}}$, respectively, and the initial non-inferiority margin, $\boldsymbol{\delta_{initial}}$.

At the final analysis, constancy will again be tested using the group-sequential procedure of Lan-Demets (1983). If $\boldsymbol{Z_t^*} > U$ when $t = 1$, where $\boldsymbol{Z_t^*}$ is calculated via equation (6) based on the complete data at the final analysis, significant non-constancy is present. Repeat the same procedure regarding the covariate-adjustment method that was implemented at the interim analysis for finding $\boldsymbol{\delta_{adjusted}}$, only with using the complete data at the final analysis. Once the adjusted non-inferiority margin is calculated, non-inferiority

is tested using either the fixed margin or synthesis methods (as pre-specified) provided by equations (2) and (3), respectively. If $\boldsymbol{Z}_t^* > L$, then non-inferiority is tested using step 6 of the TACT method.

Below are two charts summarizing the proposed adaptive two-stage method at both interim and final analyses:

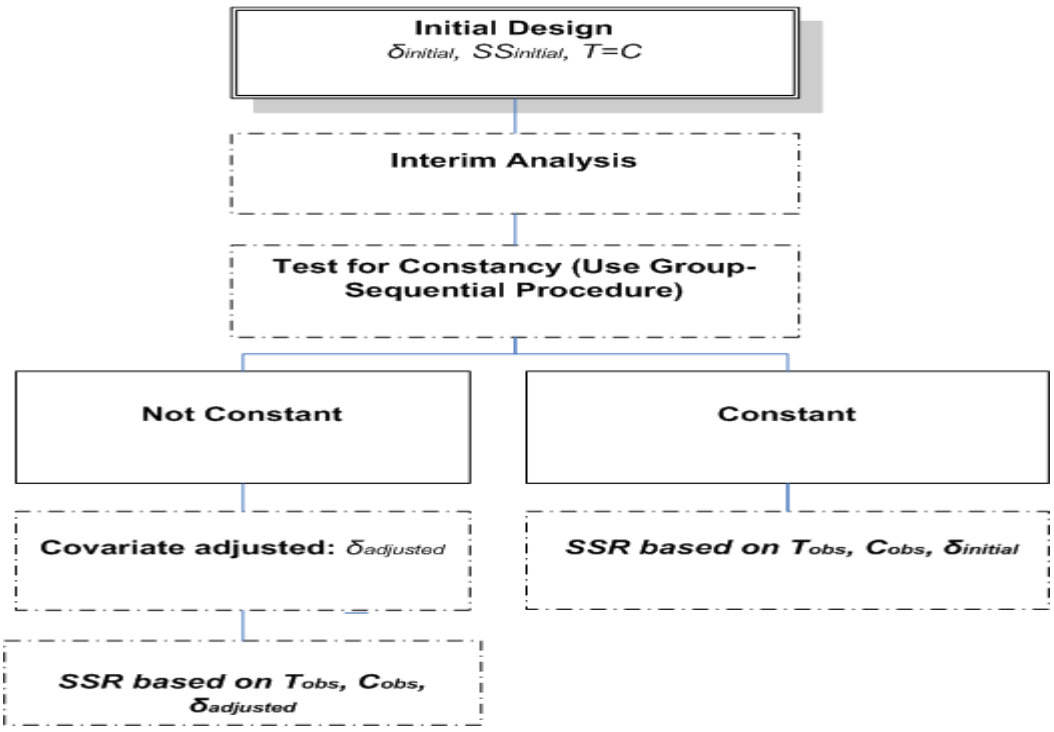


Figure 1: Proposed Adaptive Two-Stage Method (Interim Analysis)

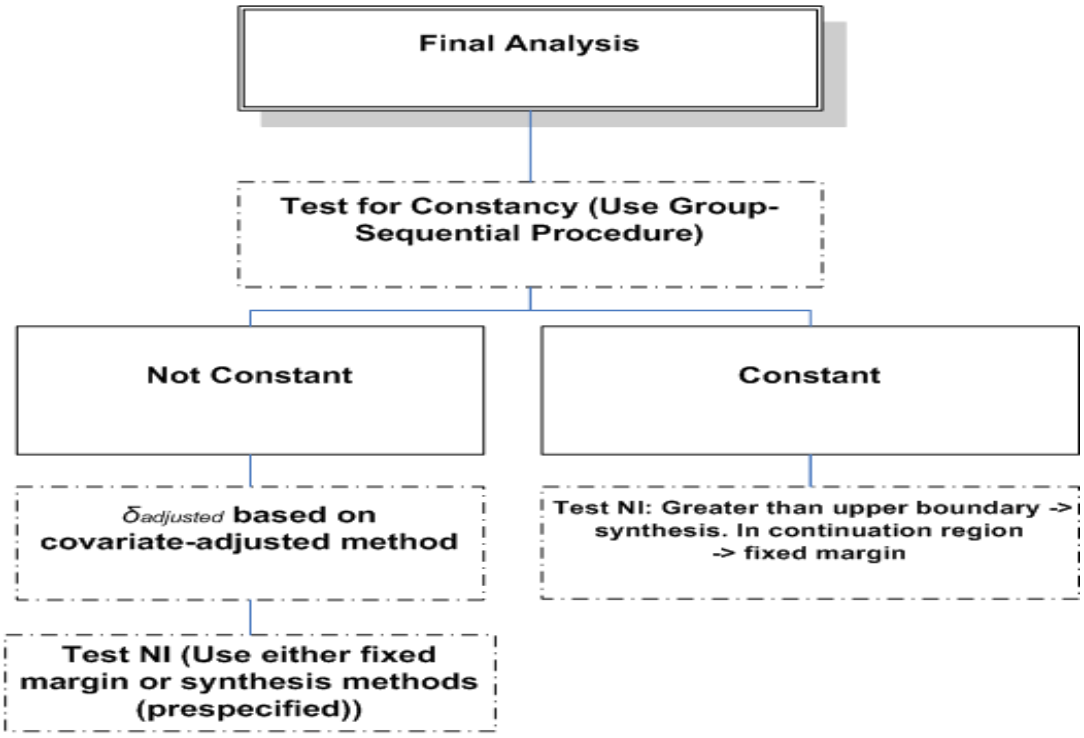


Figure 2: Proposed Adaptive Two-Stage Method (Final Analysis)

5. Simulation Studies

Simulation studies were conducted to estimate the Type I and Type II errors for the proposed 2-stage adaptive method in addition to computing the re-estimated sample size. To assess the utility of the proposed adaptive two-stage method, two simulation scenarios were implemented. The first scenario assumes $C_0 \neq C$) while the second scenario assumes $T \neq C$. The comparators are single-stage fixed margin and synthesis methods, and the maximum sample size when re-estimating the sample size will be set to be no more than double the original sample size to reflect the constraints of a realistic clinical trial setting. The non-inferiority state of truth was determined based on the synthesis method by the following equation:

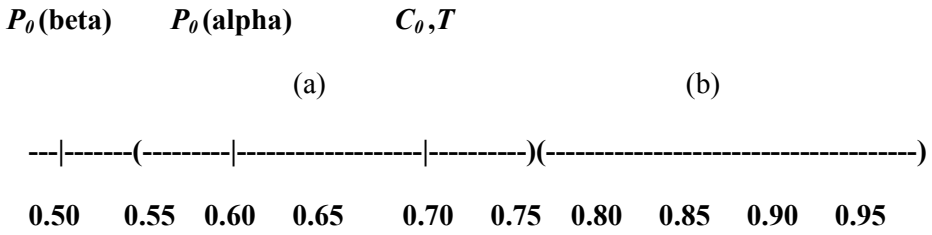
$$T\text{-}C < 0.5(P_0\text{-}\ C_0) \qquad \textbf{(9)}$$

If this inequality holds true, then the non-inferiority state of truth is satisfied. When testing for non-inferiority, the null hypothesis is that the test treatment is inferior. Hence, if equation (9) is true under a specified set of assumptions for C, T, C_0, and P_0, beta errors are reported. Likewise, if equation (9) is false, the non-inferiority state of truth is not satisfied and alpha errors are reported. Ideally, alpha error and beta error should be well-controlled; that is, $\alpha \leq 0.025$ and $\beta \leq 0.2$ for a non-inferiority trial. A total of 5000 simulation runs are generated per scenario. The set of assumptions for the two scenarios are described below:

***Scenario 1*: (Assume $C \neq C_0$)**

(a): C = 0.55 to 0.75, $C_0 = T$ = 0.7, P_0 = 0.50, η = λ = 0.5. Report beta errors

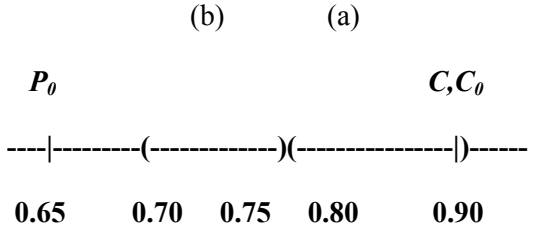
(b): C = 0.75 to 0.95, $C_0 = T$ = 0.7, P_0 = 0.60, η = λ = 0.5. Report alpha errors



***Scenario 2*: (Assume $T \neq C$)**

(a): T = 0.77 to 0.9, $C = C_0$ = 0.9, P_0 = 0.65, η = λ = 0.5. Report beta errors

(b): T = 0.70 to 0.77, $C = C_0$ = 0.9, P_0 = 0.65, η = λ = 0.5. Report alpha errors



6. Results

Figure 3 shows that under Scenario 1, both two-stage, adaptive methods always meet the target nominal levels for Type II error ($\beta \leq 0.2$, corresponding to 80% power) and outperformed both single-stage methods in terms of Type II (beta) error. Figure 4 depicts that under Scenario 1, the single-stage synthesis method performs poorly in terms of alpha error, while the other three methods are well-controlled within α = 0.025, with the two-stage adaptive fixed margin performing best. In both Figures 3 and 4, notice that when C is close to C_0, the adaptive synthesis method without covariate adjustment is implemented most often. As C either increases or decreases, moving further away from C_0, the adaptive fixed margin method without covariate adjustment is implemented more often and peaks when the absolute difference between C - C_0 is 0.1. When $|C - C_0| > 0.1$, the covariate adjustment method is implemented most frequently since this is the situation in which the constancy assumption is most violated.

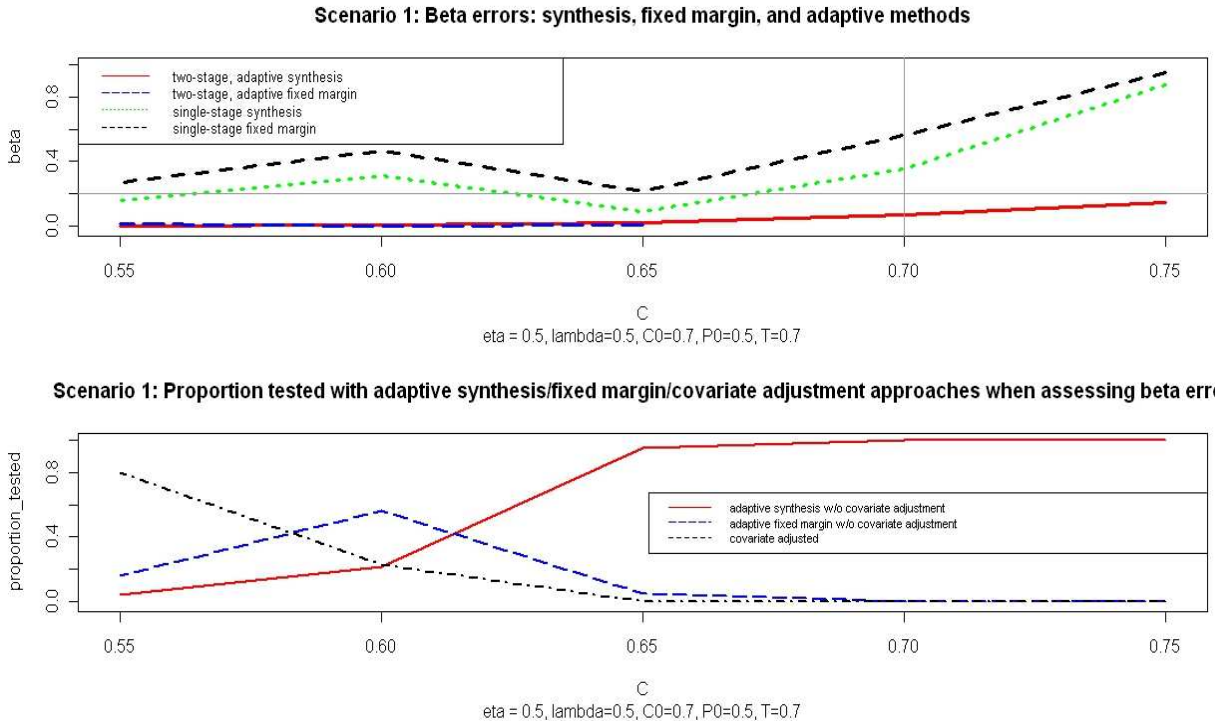


Figure 3: Beta errors and proportion tested under adaptive synthesis/fixed margin/covariate adjustment approaches ($C \neq C_0$)

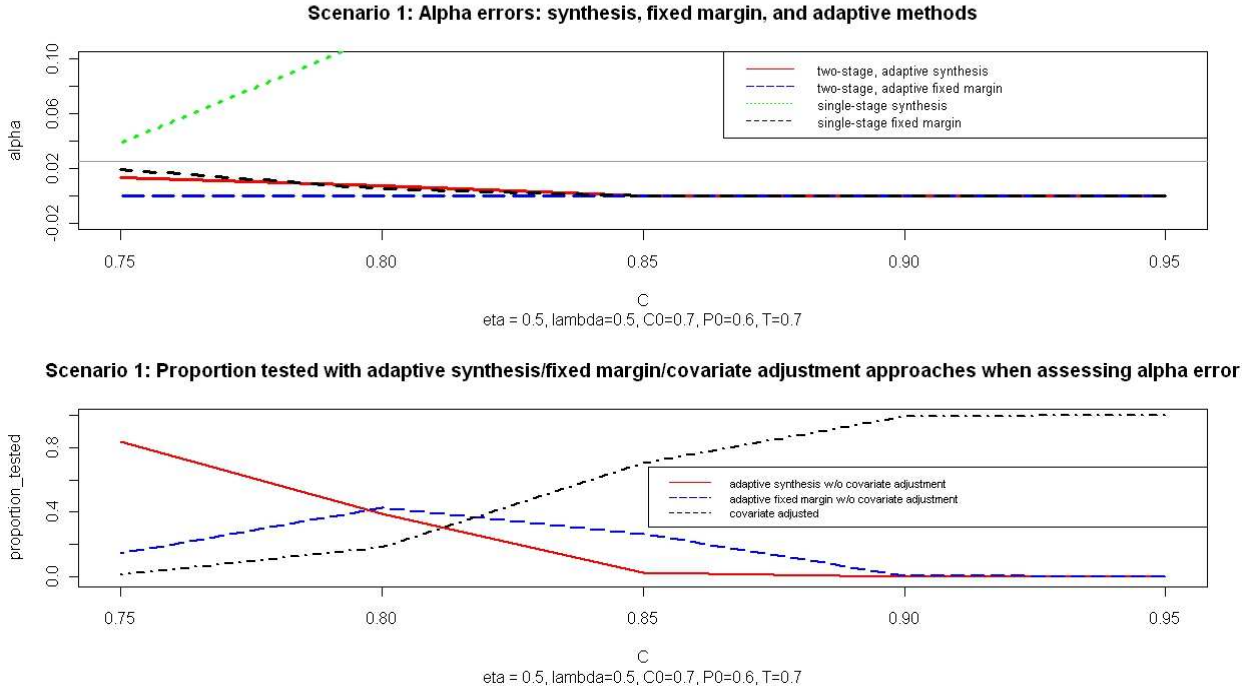


Figure 4: Alpha errors and proportion of simulation runs tested under adaptive synthesis/fixed margin/covariate adjustment approaches ($C \neq C_0$)

Figure 5 shows diagnostics regarding the proportion of simulation runs that is non-constant at the interim or final analysis. The proportion of simulation runs deemed non-constant at the interim analysis is greater when assessing alpha error than beta error. However, as C decreases and moves further away from C_0 when assessing the beta error, the number of simulation runs deemed non-constant increase precipitously as one moves from the interim to final analysis.

Figure 6 depicts the situation in which the adaptive, two-stage method is implemented without performing sample size re-estimation. From the results, the sample size re-estimation improves both alpha and beta error in the synthesis and fixed margin methods. The improvement is most pronounced regarding the two-stage adaptive synthesis method, which represents the only case where the beta error is above the desired, nominal level of 0.2 when sample size re-estimation is not implemented.

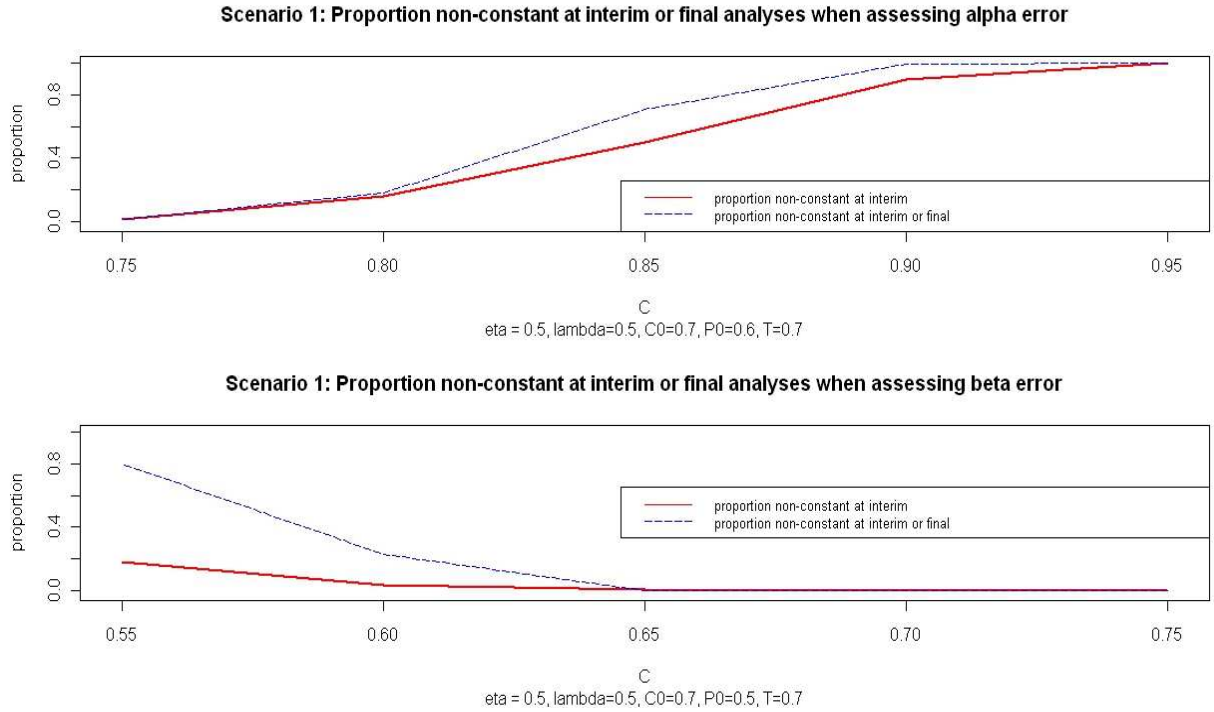


Figure 5: Proportion Non-constant at Interim or Final Analyses When Assessing Alpha or Beta Error ($C \neq C_0$)

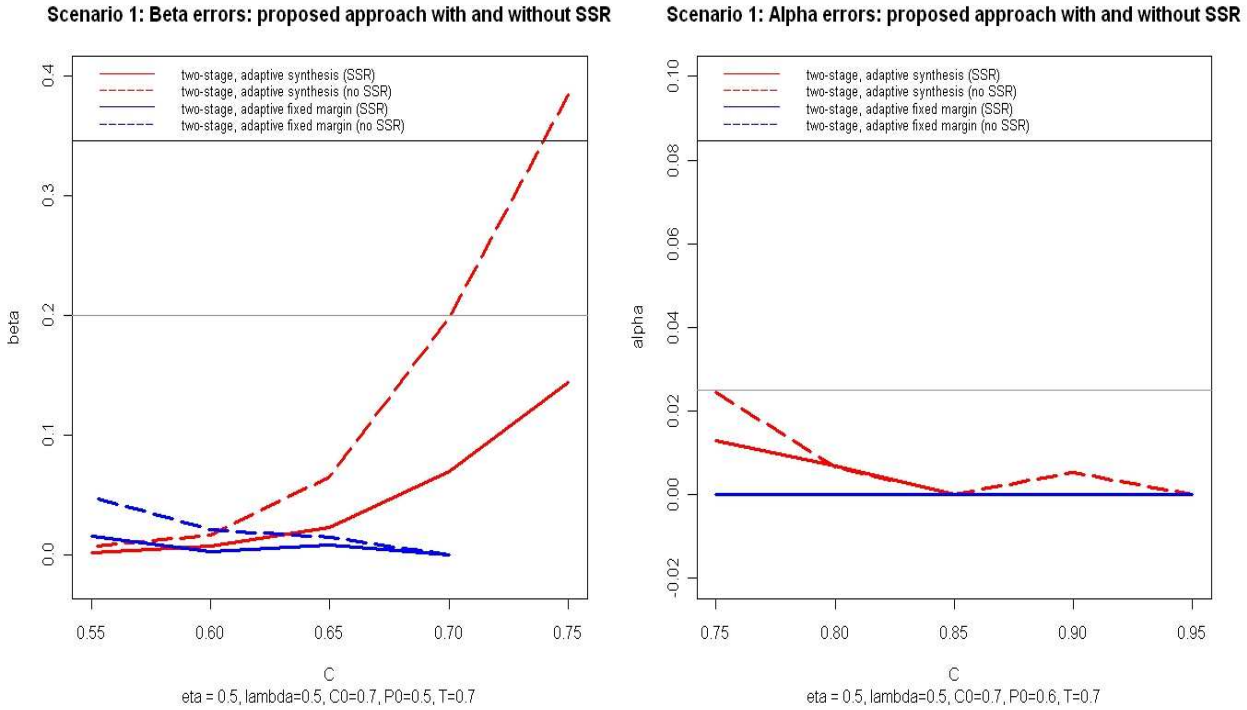


Figure 6: Alpha/beta errors under two-stage adaptive method with and without sample size re-estimation ($C \neq C_0$)

The re-estimated sample sizes are provided in Table 1. As C decreases from 0.75 to 0.55, the percentage increase in sample size (based on the median) starts at 32% but decreases to 13% with C_0 , P_0, and T remaining fixed at 0.70, 0.50, and 0.70, respectively. However, as C increases from 0.75 to 0.95, with P_0 changed to 0.60, the increase in sample size (based on the median) starts at 100% but decreases to 0%. The reason for this wide disparity of the median increase in sample size is that P_0 is very close to C_0 and T. In this case, it can also be observed that the highest median sample size required occurs when C=0.80 at n = 211. Thereafter, as C increases, the sample size levels off to around n = 135 to 165. Overall, as non-constancy between C and C_0 becomes more pronounced, the actual percent increase in subjects when performing sample size re-estimation actually decreases.

Table 1: Sample Size for Scenario 1 (C varying). $C_0 = T = 0.7$, $P_0 = 0.5$ or 0.6, $\eta = \lambda = 0.5$, C ranges between 0.55 and 0.95.

C	P_0	Original Sample Size	Re-estimated Sample Size		
			Q1	Median	Q3
0.55	0.5	144	144	211	287
0.60	0.5	161	161	213	322
0.65	0.5	171	171	207	342
0.70	0.5	172	172	198	344
0.75	0.5	167	167	188	343
0.75	0.6	92	143	184	184
0.80	0.6	106	143	211	211

0.85	0.6	119	119	151	238
0.90	0.6	135	135	135	207
0.95	0.6	165	165	165	165

For Scenario 2, Figure 7 shows that both two-stage methods (synthesis and fixed margin) exhibit less alpha and beta error than their single-stage counterparts. When T ranges from 0.70 to 0.75, the alpha error for all methods is well-controlled except for the single-stage synthesis method. When T ranges from 0.75 to 0.77, the alpha error in all methods inflate above the desired, nominal level of $\alpha = 0.025$. However, the two-stage fixed margin method overall contains the smallest alpha error and only slightly inflates above the desired, nominal level. Regarding the beta errors, each of the four methods contains a quadratic shape, which signifies that the non-inferiority effect may not be clear at the cases where the non-inferiority state of truth described by equation (9) is near the borderline for declaring either inferiority or non-inferiority. However, the two-stage synthesis method performs consistently better in terms of beta error than the other three methods.

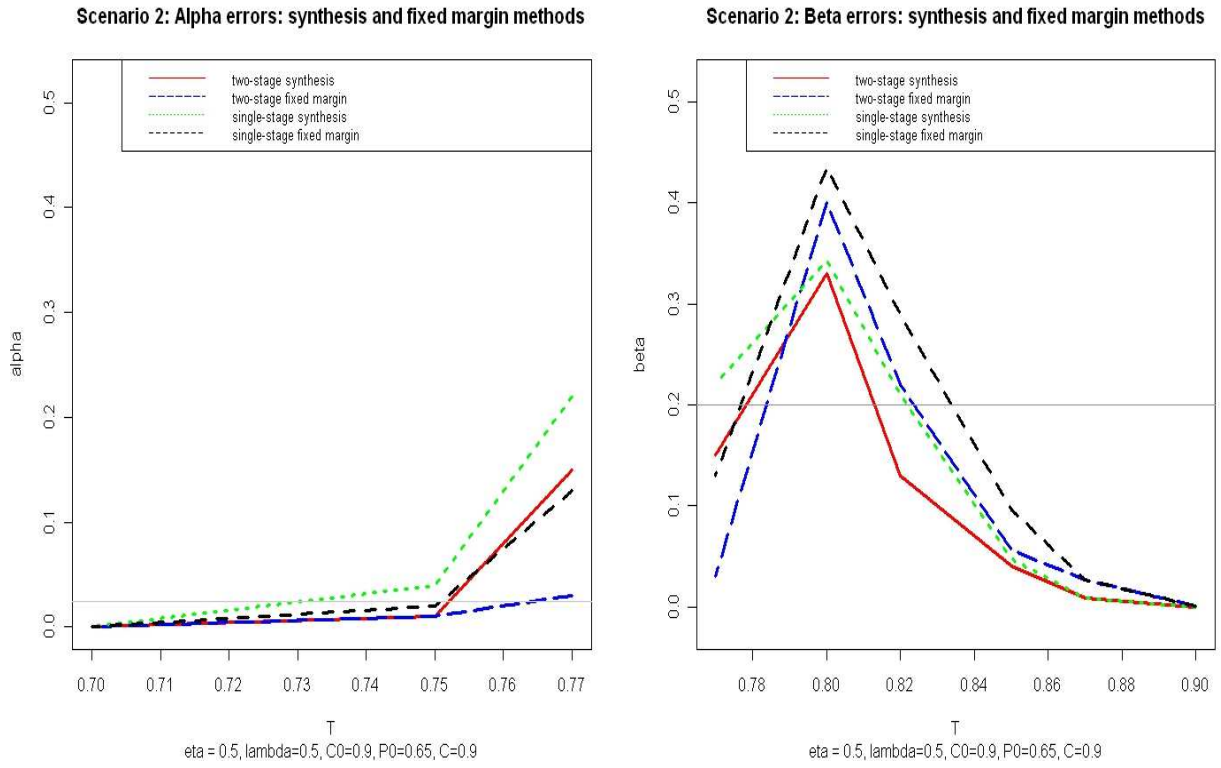


Figure 7: Alpha and beta errors for the synthesis and fixed margin methods. ($T \neq C$)

The re-estimated sample sizes are given in Table 2. When controlling for alpha error (i.e., $0.70 \leq T \leq 0.77$), the increase in sample size is narrow ranging from 16-19%. Furthermore, when controlling for beta error (i.e., $0.77 < T \leq 0.90$), the increase in sample size (based on the median) ranges from 0 to 23%. As T increases and travels further away from the fixed value of C at 0.9, there is less increase in sample size. This result is expected as the chance of detecting a difference between T and C increases.

Table 2: Sample Size for Scenario 2 (T varying). $C_0 = C = 0.9$, $P_0 = 0.65$, $\eta = \lambda = 0.5$, T ranges between 0.70 and 0.90.

C	Original Sample Size	Re-estimated Sample Size		
		Q1	Median	Q3
0.70	188	188	232	287
0.75	188	188	225	376
0.77	188	188	225	376
0.80	188	188	232	376
0.82	188	188	232	376
0.85	188	188	232	376
0.87	188	188	217	224
0.90	188	188	188	225

6. Conclusions

($C \neq C_0$)

Alpha error

1) The synthesis single-stage method led to high type I error inflation.
2) The adaptive, two-stage synthesis method prevented this inflation from occurring. The increase in sample size (based on the median) ranges from 0-100% (dependent on the difference of $P_0 - C$).

Beta error

1) The beta error was inflated for both single-stage methods (worse for fixed margin).
2) Both adaptive, two-stage methods alleviated this problem. The increase in sample size (based on the median) ranges from 13-32%.
3) Sample size re-estimation had the greatest impact for reducing the beta error within the synthesis method.

($T \neq C$)

Alpha error

1) SSR decreased alpha in the adaptive, two-stage methods as compared to the respective single-stage methods.
2) As T moves further away from C, the alpha error was well-controlled in both two-stage methods and the single-stage fixed margin method. The increase in sample size (based on the median) ranges from 16-19%.

Beta error

1) The two-stage methods had an overall smaller beta error than the respective single-stage methods.
2) The two-stage synthesis method had the lowest overall beta error. The increase in sample size (based on the median) ranges from 0-23%.

References

1) CBER and CDER FDA Memorandum. Guidance for Industry: Non-Inferiority Clinical Trials, March 2010.
2) D'agostino, Massaro, and Sullivan. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**: 169-186.
3) Lan KKG, DeMets DL. Design and analysis of group sequential tests based on the type I error spending function. *Biometrika* 1983; **74**: 149-154
4) Nie, L and Soon G. A covariate-adjustment regression model approach to noninferiority margin definition. *Statistics in Medicine* 2010; **29**: 1107-1113.
5) Rothmann M, Wiens B, Chan I. Design and Analysis of Non-Inferiority Clinical Trials, CRC Press, Boca Raton, FL 2012; **5**: 117-120.
6) Simon R. Bayesian design and analysis of active controlled clinical trials. *Biometrics* 1999; **55***:* 484-487.
7) Wang, SJ and Hung, HMJ. TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**: 227-238.
8) Wang SJ, Hung HMJ, Tsong Y. Utility and pitfall of some statistical methods in active controlled clinical trials. *Controlled Clinical Trials* 2002; **23**(1):15-28.