Bayesian Augmented Control Methods for Efficiently Incorporating Historical Information in Clinical Trials

Carl DiCasoli[1], Michael Kunz[1], Daniel Haverstock[1]
[1]Bayer Healthcare Pharmaceuticals, Whippany, NJ and Berlin, Germany

Abstract

When planning a clinical trial, there is often historical clinical data available. Recently Viele *et al.* (2013) have presented approaches that incorporate this historical clinical data into an analysis procedure. We focus on the idea of dynamic borrowing in the framework of various hierarchical modeling strategies and derive the Type I error, power, and DIC. These strategies may include estimating parameters for each trial separately versus pooling, weighting the prior distribution corresponding to each historical study based on the sample size, and incorporating historical borrowing on the control arm by a separate random effect parameter. As a further refinement we present how in the setting of a non-inferiority trial a covariate adjustment approach can be implemented to recalibrate the non-inferiority margin based on the difference between active control and placebo.

Key Words: Bayesian, hierarchical, noninferiority, Type I error, power, covariate adjustment

1. Introduction

In late-phase (Phase III/IV) clinical trials, a primary primary objective was to establish non-inferiority of a test treatment, *T*, versus an active control treatment, *C*, with respect a binary outcome variable, where high values 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., $C\text{-}P \approx C_0\text{-}P_0$). If this constancy assumption is violated, new methodology via a covariate-adjustment model could salvage the active control trial. T, C, P, C_0, P_0, represent event rates associated with the experimental treatment, active control treatment, active placebo, historical control treatment, and historical placebo, respectively

Current frequentist methodology regarding non-inferiority constitutes four methods, three of which are standard: testing the difference between treatment and active control when no placebo arm is available (referred to as the "*T*-*C*" method, the conservative confidence interval (or fixed margin method), and the synthesis method. The fourth frequentist non-inferiority method is the covariate-adjustment model of Nie and Soon (2010) to address non-constancy ($C_0 \neq C$) arising from heterogeneity between patient populations in the active and historical trials.

However, prior to utilizing these frequentist methods, a model is initially fit that already incorporates historical borrowing on the control arm via a Bayesian hierarchical approach. Some authors that have examined Bayesian approaches in past literature include Pennello and Thompson (2008) who used a

hierarchical model to borrow strength from historical controls to reduce the sample size of the active control and Ibrahim and Chen (2000) who examined covariates in power prior models, where a power prior is derived by raising the historical likelihood to a particular power. More recently, the general idea has been examined by Viele *et al.* (2013) through an augmented hierarchical method that effectively shares more information when the information between active and historical trials is similar and less information when the information between active and historical trials is disparate. These authors explained that in certain situations calibrating the active and historical controls is necessary to arrive at the correct inferences. The approach that will be outlined below not only implements the Bayesian model to more effectively share information when the active and historical trials are homogeneous, but follows with further, frequentist approaches, one of which can potentially salvage the active control trial if substantial heterogeneity is present. Hence, this approach can adapt to many possible situations that arise from the reliability of previous historical trial information.

2. The Four Non-Inferiority Frequentist Inference Methods

2.1 Testing the Difference between Treatment and Active Control

The strategy of this non-inferiority inference method is to declare the trial a success if:

$$\mathbf{Pr}_{H_0}(\widehat{T} > \widehat{C} - \boldsymbol{\delta}_{\mathbf{initial}}) < 0.025, \qquad (1)$$

where $\boldsymbol{\delta}_{\mathbf{initial}}$ represents the non-inferiority margin; that is, the portion of effectiveness of active controlthat may be lost in the performance of the test drug, and is known in non-inferiority literature as M2, whereas M1 is referred to as the lower bound of historical treatment effect of active control drug versus placebo. After standardizing, a test statistic can be expressed as:

$$Z_{ni} = [T - C + \boldsymbol{\delta}_{\mathbf{initial}}] / [\boldsymbol{\sigma}_{TC}], \qquad (2)$$

$\boldsymbol{\sigma}_{TC}$ represents the standard error of $T - C$ in the active trial. This method is usually implemented in many non-inferiority studies that do not contain placebo arms. Hence, $\boldsymbol{\delta}_{\mathbf{initial}}$ in this particular approach does not take into account the difference between $\Pr(P_0 = 1)$ and $\Pr(C_0 = 1)$.

2.2 The Fixed Margin and Synthesis Methods

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 $\boldsymbol{\delta}_{\mathbf{initial}}$, as defined above. In the context of the fixed margin method, we can notate specifically,

$$\mathbf{M2} = (1-\boldsymbol{\eta})\,\mathbf{M1}, \qquad (3)$$

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

At the analysis stage, it is required to show that the upper bound of the 95% confidence interval for $\Pr(T=1) - \Pr(C=1)$ and T is within the specified non-inferiority margin, $\boldsymbol{\delta}_{\mathbf{initial}}$. That is,

$$T - C + z_{0.025}\,\sigma_{TC} < (1-\eta)\{P_0 - C_0 - z_{0.025}\,\sigma_{P0C0}\}, \qquad (4)$$

where σ_{TC} represents the standard deviation of $T - C$ 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 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

$$Z_{pv} = [T - C - (1-\lambda)\{P_0 - C_0\}] \,/\, [\mathrm{sqrt}(\sigma^2_{TC} + (1-\lambda)^2\,\sigma^2_{P0C0})], \qquad (5)$$

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

2.3 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 while still 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**:

$$E(y_i) = g^{-1}(\mu_i),\ \mu_i = \alpha + \beta Z_i + \sum_{k=1}^{K}(\beta_k x_{ik} + \gamma_k x_{ik} Z_i), \qquad (6)$$

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

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 β will change with the covariates x_{ik}.

If the constancy assumption is rejected, the NI margin $\delta_{adjusted}$ is recalibrated to the active control population via the regression model in (6) and is defined as the lower bound of a (1-α)100 % CI of $P–C$, where

$$P - C = \beta + \sum_{k=1}^{K} \gamma_k\, \bar{x}_{.k}, \qquad (7)$$

and $\bar{x}_{.k}$ represents the mean of the active control population, $\hat{C}$. The calibrated estimate of $\delta_{adjusted}$ 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 (6). This covariate adjustment can be implemented for the fixed margin and 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 *T*-*C* 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 (*T*-*C*)-(1-λ)(*C*-*P*) < 0 on the transformed scale of choice.

3. Bayesian Augmented Control Methods

The covariate-adjusted method of Nie and Soon (2010) employs an underlying frequentist regression model before performing covariate adjustment, while both the fixed margin and synthesis methods have notable drawbacks. The frequentist method described in the previous section (synthesis, fixed margin, and covariate adjustment) could be improved via more accurate point estimates, increased power and reduced Type I error if the initial model fit to obtain estimates and distributions of $\boldsymbol{T}$, $\boldsymbol{C}$, $\boldsymbol{C_0}$,$\boldsymbol{P_0}$ incorporated historical borrowing on the control arm via a Bayesian, hierarchical model rather than a frequentist, general linear model (GLM) that does not incorporate this additional learning.

If the information between $\boldsymbol{C}$ and $\boldsymbol{C_0}$ is similar, more will be borrowed. If not, less information will be borrowed. Hence, this may prove to be an advantage for those cases with moderate to high similarity in the rates between $\boldsymbol{C}$ and $\boldsymbol{C_0}$. In those cases were $\boldsymbol{C}$ and $\boldsymbol{C_0}$ are not similar, the properties of point estimate accuracy, increased power, and Type I error could possibly show further improvement after fitting the Bayesian hierarchical model by subsequently performing the covariate adjustment method via recalibration of the non-inferiority margin based on available additional information, e.g., suppose that patients in a particularly sick subgroup that constitute a substantial portion of the population (~15%) experience success rates of response at half of those in the less sick subgroup. The four proposed Bayesian hierarchical models along with their corresponding assumptions are outlined below in Table 1.

Table 1: The Four Bayesian Augmented Hierarchical Models

Name of Model	Model Statement	Priors	Assumptions
(1) Hierarchical borrowing, separate historical trials	$\log(\boldsymbol{p}/(\boldsymbol{1-p})) = \boldsymbol{\beta_1 I_{\{C=1 \text{ or } T=1\}} + \beta_2 I_{\{C01=1\}} + \beta_3 I_{\{C02=1\}} + \delta I_{\{P0=1\}} + \theta I_{\{T=1\}}}$	$\boldsymbol{\beta_k \sim}$ **Normal**$(\boldsymbol{\mu}, \boldsymbol{\omega})$, $\boldsymbol{\delta \sim}$ **Normal**$(0, 10^{-6})$, $\boldsymbol{\theta \sim}$ **Normal**$(0, 10^{-6})$, $\boldsymbol{\mu \sim}$ **Normal** $(1, 1)$, $\boldsymbol{\omega^2 \sim}$ **InvGamma**$(0.01, 0.01)$, $\boldsymbol{k = 1, 2, 3}$. ***The second parameter represents precision (throughout all models).***	$\boldsymbol{C_{01}}$ **and** $\boldsymbol{C_{02}}$ represent the historical control effect for trials 1 and 2, respectively, **ω** represents the across study precision which inversely corresponds to the degree of borrowing (greater **ω** = more borrowing), $\boldsymbol{p}$ represents the success proportion of each group; i.e., $\boldsymbol{C}$, $\boldsymbol{C_{01}}$, $\boldsymbol{C_{02}}$, $\boldsymbol{P_0}$, $\boldsymbol{T}$

(2) Hierarchical borrowing, weighed sharing effect	$\log(p/(1-p)) = \beta_1 I_{\{C=1 \text{ or } T=1\}} + \beta_2 I_{\{C01=1\}} + \beta_3 I_{\{C02=1\}} + \delta I_{\{P0=1\}} + \theta I_{\{T=1\}}$	$\beta_1 \sim$ **Normal**(μ, ω_1), $\beta_2 \sim$ **Normal**$(\mu, \omega_2 = \omega_1 \times k_1^2)$, $\beta_3 \sim$ **Normal**$(\mu, \omega_3 = \omega_1 \times k_2^2)$, $\delta \sim$ **Normal**$(0, 10^{-6})$, $\theta \sim$ **Normal**$(0, 10^{-6})$, $\mu \sim$ **Normal**$(1, 1)$, $\omega_k^2 \sim$ **InvGamma(** 0.01, 0.01)	k_m represents the factor that links the sample size of C with C_{0m}, for m = 1, 2. (e.g., n_c = 200 but n_{C01} = 100, -> k_1 = 0.5)
(3) Hierarchical borrowing, pooled historical trials	$\log(p/(1-p)) = \beta_{1\{C=1 \text{ or } T=1\}} + \beta_{2\{C0p=1\}} + \delta I_{\{P0=1\}} + \theta I_{\{T=1\}}$,	$\beta_k \sim$ **Normal**(μ, ω), $\delta \sim$ **Normal**$(0, 10^{-6})$, $\theta \sim$ **Normal**$(0, 10^{-6})$, $\mu \sim$ **Normal** $(1, 1)$, $\omega \sim$ **InvGamma(** 0.01, 0.01)., k **= 1, 2**	C_{0p} represents the overall historical control effect pooled across all available trials.
(4) Hierarchical borrowing, free-floating sharing	$\log(p/(1-p)) = \beta_{1\{C=1 \text{ or } T=1\}} + \beta_{2\{C0p=1\}} + \delta I_{\{P0=1\}} + \theta I_{\{T=1\}} + \psi_r$	$\beta_k \sim$ **Normal**$(\beta_k^*, 10^{-6})$, $\delta \sim$ **Normal**$(\delta^*, 10^{-6})$, $\theta \sim$ **Normal**$(\theta^*, 10^{-6})$, β^*, θ^*, $\delta^* \sim$ **Normal(**$0, 10^{-6}$), $\psi_q^* \sim$ **Normal**$(0, \omega^2)$, $\omega \sim$ **Half-normal**(1), k**= 1, 2,** r**=1,2,3**	p = 1, 2 represents random effects placed on C and C_{0p}, r = 3 represents random effects pooled across all other arms **(**T**,** P_0**,)**

The purpose of implementing the first two models "Hierarchical borrowing, separate historical trials" and "Hierarchical borrowing, weighted sharing effect" was to compare the Bayesian hierarchical model when the prior distribution is weighted according to the sample size (more information is placed when the sample size is larger) versus four other borrowing strategies between C and C_0. In summary,

1. **Hierarchical borrowing (weighted).** This situation corresponds to the situation outlined in Model #2; that is, more borrowing from the historical or active control arm that contain the higher sample size.
2. **Full borrowing.** The precision (1/variance) for **ω** within the individual priors for β_k, k = 1,2,3 is set at **ω**=1000 (most informative) to borrow the maximum 200 subjects.
3. **Half borrowing.** The precision (1/variance) for **ω** within the individual priors for β_k is set at **ω**=100 (informative) to borrow 100 subjects.
4. **Hierarchical borrowing (unweighted).** This situation corresponds to the situation outlined in Model #1.
5. **No borrowing.** The precision (1/variance) for **ω** within the individual priors for β_k is set at **ω**=10^{-6} (least informative) to borrow approximately 0 subjects.

The number borrowed was calculated according to the formula

$$\frac{(\hat{p}_c(1-\hat{p}_c))}{\hat{\sigma}^2_{C0(average)}} - n_c, \qquad \textbf{(8)}$$

where $\hat{p}_c$ represents the proportion of success in the active control arm, $\hat{\sigma}^2_{C0(average)}$ represents the squared standard error of the average of the proportion of success across the two separate historical control arms, and n_c represents the sample size of the active control arm.

The strategy regarding the third model (hierarchical borrowing, pooled historical trials) is similar to that of the "Hierarchical borrowing, separate historical trials" model; that is, the precision within the priors for

$\boldsymbol{\beta_k}$ does not depend upon the individual sample sizes of the historical and active control arms. The main difference between these two models is that in the data from all historical control arms across different trials (assumed to be the base case of two trials shown in Table 1) are "pooled" into one historical control parameter, $\boldsymbol{C_{0p}}$, instead of fitting each trial parameter regarding the historical control arm separately. The following scenarios were studied:

1. "Hierarchical borrowing, separate historical trials", $\boldsymbol{C_{01}}$ = 0.68, $\boldsymbol{C_{02}}$ = 0.72.
2. "Hierarchical borrowing, separate historical trials", $\boldsymbol{C_{01}}$ = 0.65, $\boldsymbol{C_{02}}$ = 0.75.
3. "Hierarchical borrowing, separate historical trials", $\boldsymbol{C_{01}}$ = 0.60, $\boldsymbol{C_{02}}$ = 0.80.
4. "Hierarchical borrowing, pooled historical trials": generate simulated data from $\boldsymbol{C_{01}}$ = 0.60, $\boldsymbol{C_{02}}$ = 0.80, but fit a model with a "pooled" $\boldsymbol{C_{0p}}$=0.70.

The purpose is two-fold: 1. to compare the result of having the average rate of two trials (within trials) = 0.70, but with increasing the difference in rates between trials and 2. to compare the situation with fitting a model that specifies each individual trial separately versus a model that pools the information into one parameter when there is a large difference in the underlying active control rates from trial to trial; e.g., $\boldsymbol{C_{01}}$ = 0.60 but $\boldsymbol{C_{02}}$ = 0.80.

Finally, the fourth model (hierarchical borrowing, free-floating sharing) implements a technique different than the other three previous models. This model is similar to the approach of Jones *et al.,* (2011) who fit a "simple shrinkage" model that assumed full exchangeability within the subgroup setting. Here, the parameter **ω,** which measures the strength of historical borrowing, is free-floating rather than embedded hierarchically within the active and historical control arms as described in the other models. This approach in the context of non-inferiority trials could be useful if little previous historical information is known regarding the control arms and could potentially be of great benefit to avoid spurious conclusions.

4. Simulation Studies

Simulation studies were conducted to estimate the Type I and Type II errors, in addition to the number borrowed, updated non-inferiority margin after covariate-adjustment method, bias (relative bias and bias ratio), and DIC for the four proposed methods that were explained in Section 3.

When conducting the simulation, 100 simulation runs were implemented with unequal sample sizes; e.g., $n_{T=}$ n_C =200, while n_{C01}= n_{C02} = n_{P0} = 100, for a total of 700 subjects. For the weighted scenario, since n_C = 200 but n_{C01} = 100 and n_{C02} = 100, -> $\boldsymbol{k_1 = k_2}$ = 0.5. The initial non-inferiority margin, prior to implementing the covariate-adjustment method will be set at $\boldsymbol{\delta_{initial}}$ = 0.10. To utilize the covariate-adjustment method, assume that patients in a particularly sick subgroup experience success rates at half of those in the less sick subgroup. Additionally, patients in this subgroup constitute approximately 15% of the population. The "truth" values for each of the five arms are as follows: $\boldsymbol{C}$ = 0.55 to 0.95**,** $\boldsymbol{C_{01}}$ = 0.68**,** $\boldsymbol{C_{02}}$ = 0.72**,** $\boldsymbol{P_0}$ = 0.50, $\boldsymbol{T}$ = 0.70. Note that the pooled data between the two historical control arms, $\boldsymbol{C_{01}}$ and $\boldsymbol{C_{02}}$,with equal sample sizes is 0.70; that is, let $\boldsymbol{C_{0p}}$=0.70. In the case of Models 3 and 4, $\boldsymbol{C_{01}}$ and $\boldsymbol{C_{02}}$ are simulated from the scenarios described in Section 3; that is, 1. $\boldsymbol{C_{01}}$ = 0.68, $\boldsymbol{C_{02}}$ = 0.72, 2. $\boldsymbol{C_{01}}$ = 0.65, $\boldsymbol{C_{02}}$ = 0.75, 3. $\boldsymbol{C_{01}}$ = 0.60, $\boldsymbol{C_{02}}$ = 0.80, and 4. fit a model with a "pooled" $\boldsymbol{C_{0p}}$=0.70, but generate simulated data from $\boldsymbol{C_{01}}$ = 0.60, $\boldsymbol{C_{02}}$ = 0.80.

In our hypothesis for testing non-inferiority it is assumed that: H_0: inferior, H_A: non-inferior. Hence, Type I error -> declaring ***T*** non-inferior when ***T*** is inferior while Type II error -> declaring ***T*** inferior when ***T*** is non-inferior. 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. The set of assumptions for the two scenarios are described below:

(Assume $C \neq C_0$)

(a): C = 0.55 to 0.75, $C_0 = T = 0.7$, $P_0 = 0.50$, $\eta = \lambda = 0.5$. Report beta errors
(b): C = 0.80 to 0.95, $C_0 = T = 0.7$, $P_0 = 0.50$, $\eta = \lambda = 0.5$. Report alpha errors.

P₀ ***C₀ ,T***

(a) (b)

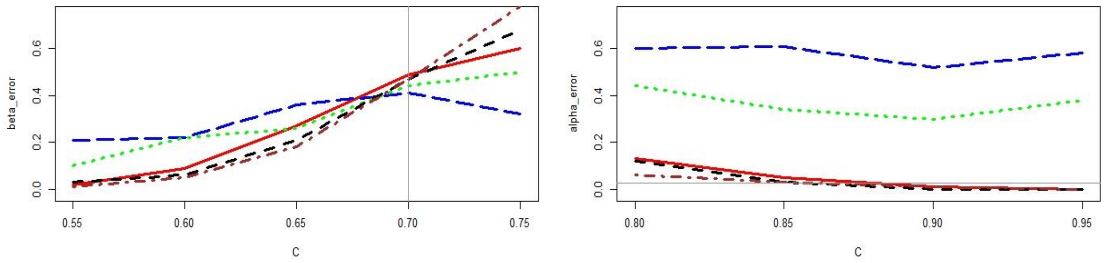
---|-------(----------|---------)(----------)

0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95

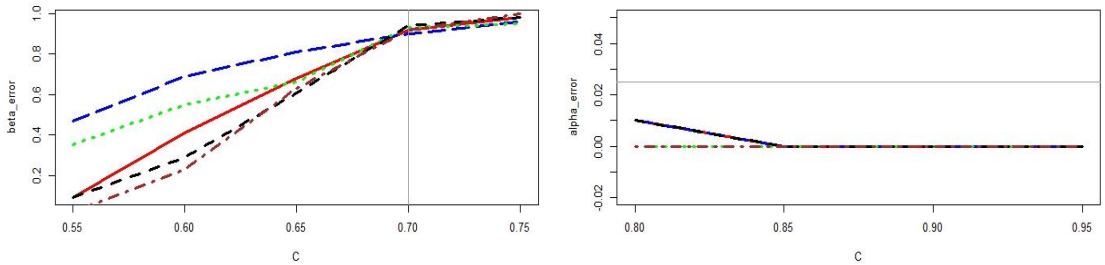
5. Results

(1) **(2)**

(1)



(2)



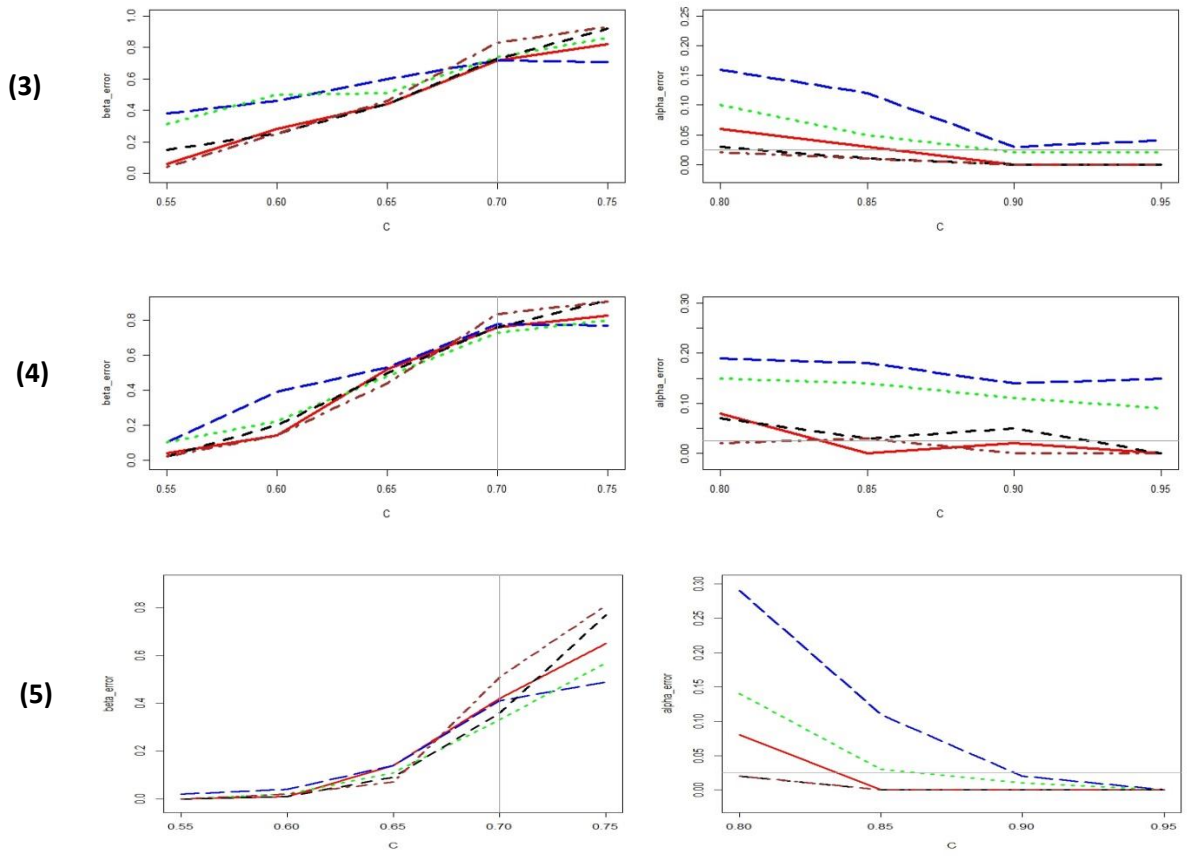


Figure 1: Comparison of the models "Hierarchical borrowing, separate historical trials" versus "Hierarchical borrowing, weighed sharing effect" for each of the borrowing scenarios symbolized by red (hierarchical borrowing (weighted)), blue (full borrowing), green (half borrowing), black (hierarchical borrowing (unweighted)), and brown (no borrowing). (1, 1) = "Baseline Synthesis Beta Errors", (1,2) = "Baseline Synthesis Alpha Errors", (2,1) = "Baseline Fixed Margin Beta Errors", (2,2) = "Baseline Fixed Margin Alpha Errors", (3,1) = "Covariate-Adjusted Synthesis Beta Errors", (3,2) = "Covariate-Adjusted Synthesis Alpha Errors", (4,1) = "Covariate-Adjusted Fixed Margin Beta Errors", (4,2) = "Covariate-Adjusted Fixed Margin Alpha Errors", (5,1) = "***T-C*** Beta Errors", (5,2) = "***T-C*** Alpha Errors". The horizontal line corresponds to $\boldsymbol{a}$=0.025, and vertical line corresponds to $\boldsymbol{C_0}$ = 0.7. y-axis = alpha or beta error, x-axis = active control "***C***"

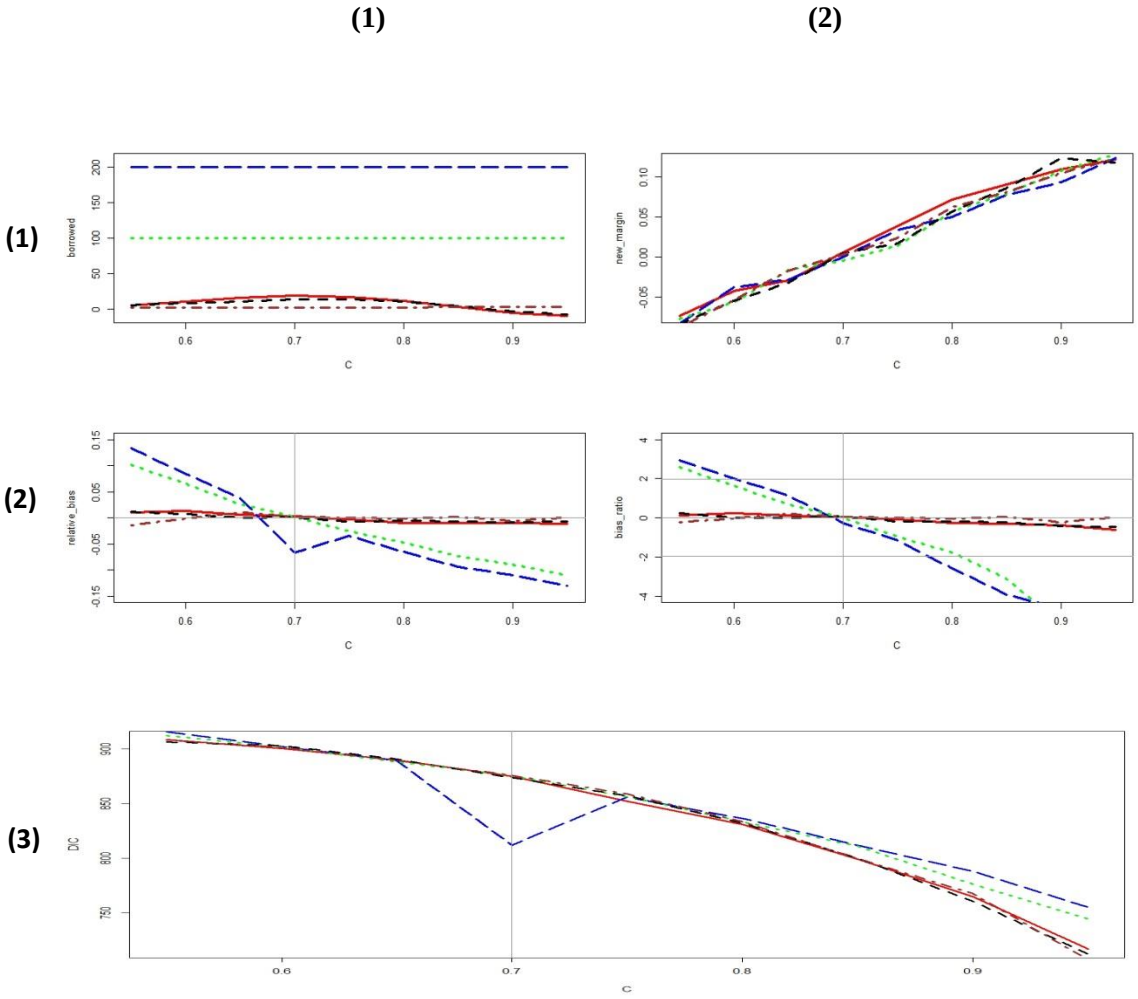


Figure 2: Model diagnostics for comparing "Hierarchical borrowing, separate historical trials" versus "Hierarchical borrowing, weighed sharing effect" for each of the borrowing scenarios listed above in Figure 1. Legend remains the same as Figure 1; that is, (hierarchical borrowing (weighted)), blue (full borrowing), green (half borrowing), black (hierarchical borrowing (unweighted)), and brown (no borrowing). (1, 1) = "Number Borrowed", (1,2) = "Adjusted, New Margin after Covariate Adjustment", (2,1) = "Relative Bias", (2,2) = "Bias Ratio", (3, ·) = "DIC." Horizontal lines correspond to $\boldsymbol{Z} = \pm 1.96$, vertical line corresponds to $\boldsymbol{C_0} = 0.7$. y-axis = diagnostic of interest, x-axis="$\boldsymbol{C}$" (active control)

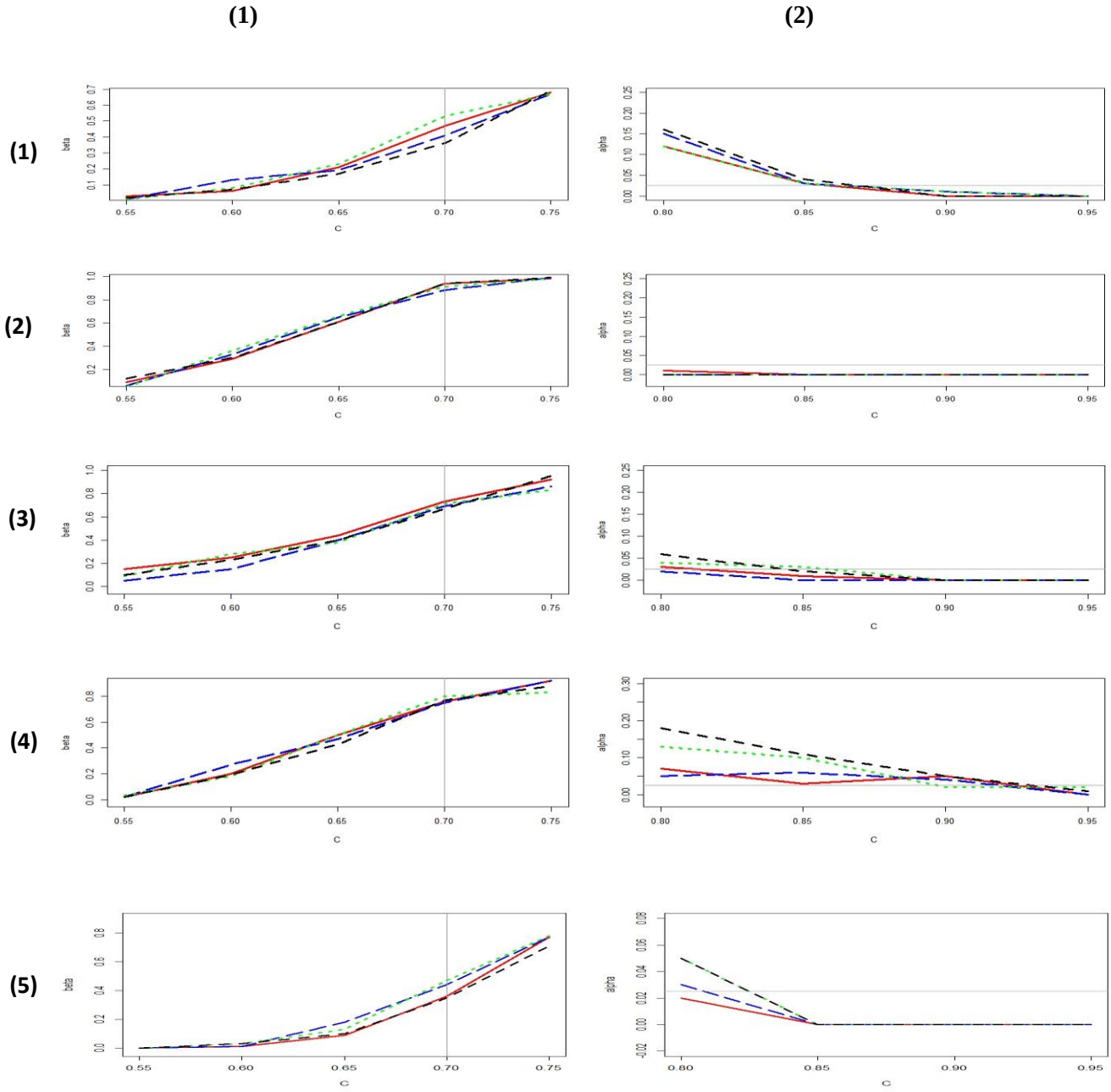


Figure 3: Comparison of the models "Hierarchical borrowing, separate historical trials" versus "Hierarchical borrowing, pooled historical trials" for each of the scenarios symbolized by red ("C_{01} = 0.68, C_{02} = 0.72, separate historical trials"), blue ("C_{01} = 0.65, C_{02} = 0.75, separate historical trials"), green ("C_{01} = 0.60, C_{02} = 0.80, separate historical trials"), and black ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials"). (1, 1) = "Baseline Synthesis Beta Errors", (1,2) = "Baseline Synthesis Alpha Errors", (2,1) = "Baseline Fixed Margin Beta Errors", (2,2) = "Baseline Fixed Margin Alpha Errors", (3,1) = "Covariate-Adjusted Synthesis Beta Errors", (3,2) = "Covariate-Adjusted Synthesis Alpha Errors", (4,1) = "Covariate-Adjusted Fixed Margin Beta Errors", (4,2) = "Covariate-Adjusted Fixed Margin Alpha Errors", (5,1) = "*T*-*C* Beta Errors", (5,2) = "*T*-*C* Alpha Errors". The horizontal line corresponds to a=0.025, and vertical line corresponds to C_0 = 0.7. y-axis = beta/alpha error. x-axis ="*C*" (active control).

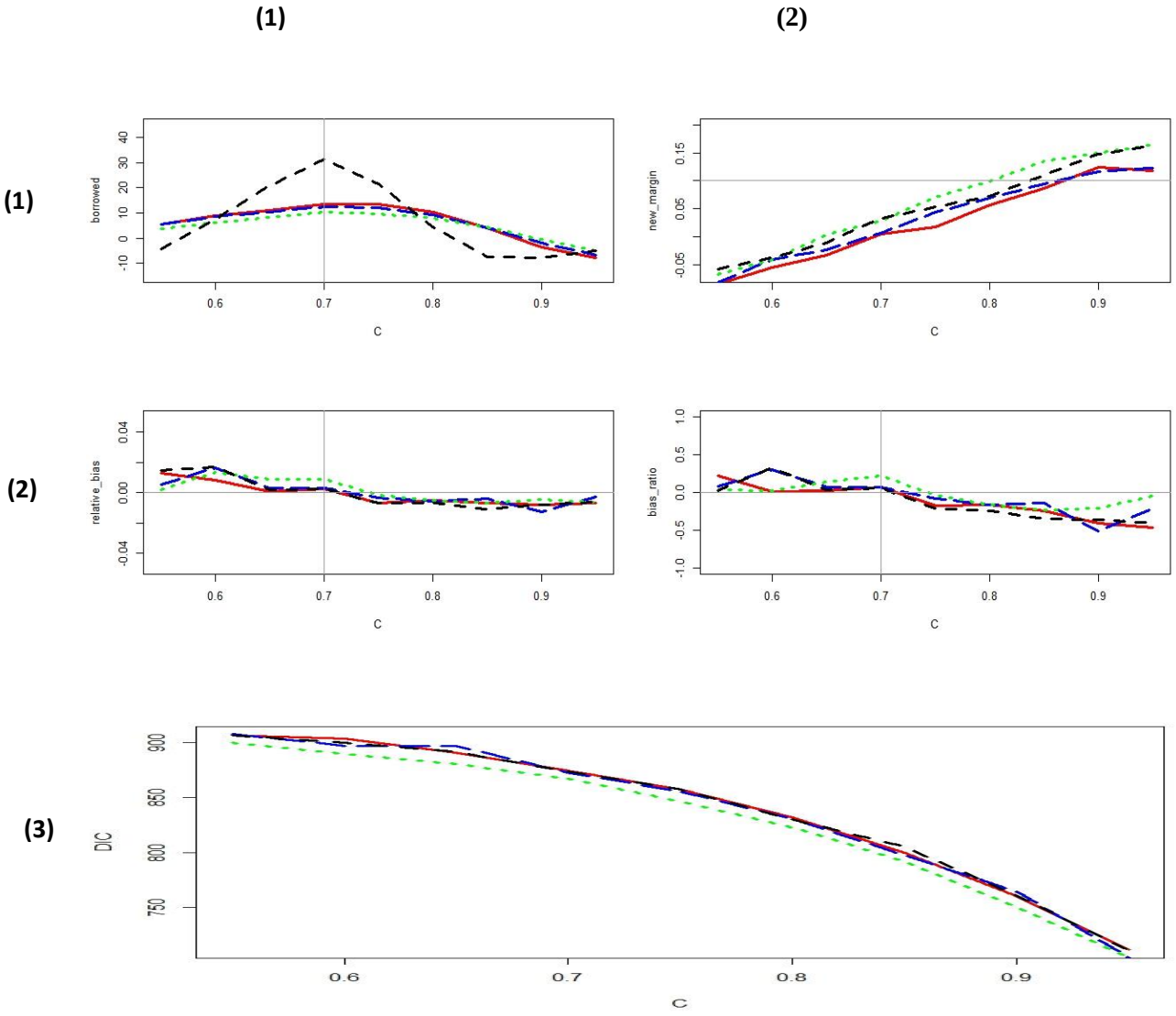


Figure 4: Model diagnostics for comparing "Hierarchical borrowing, separate historical trials" versus "Hierarchical borrowing, pooled historical trials" for each of the scenarios listed above in Figure 3. Legend remains the same as Figure 3; that is, red ("C_{01} = 0.68, C_{02} = 0.72, separate historical trials"), blue ("C_{01} = 0.65, C_{02} = 0.75, separate historical trials"), green ("C_{01} = 0.60, C_{02} = 0.80, separate historical trials"), and black ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials"). (1, 1) = "Number Borrowed", (1,2) = "Adjusted, New Margin after Covariate Adjustment", (2,1) = "Relative Bias", (2,2) = "Bias Ratio", (3,·) = "DIC." Horizontal lines correspond to $Z = 0$ (relative bias and bias ratio), $\delta_{adjusted}$ = 0.10 (new margin), vertical line corresponds to $C_0 = 0.7$. y-axis = diagnostic of interest, x-axis = "C" (active control).

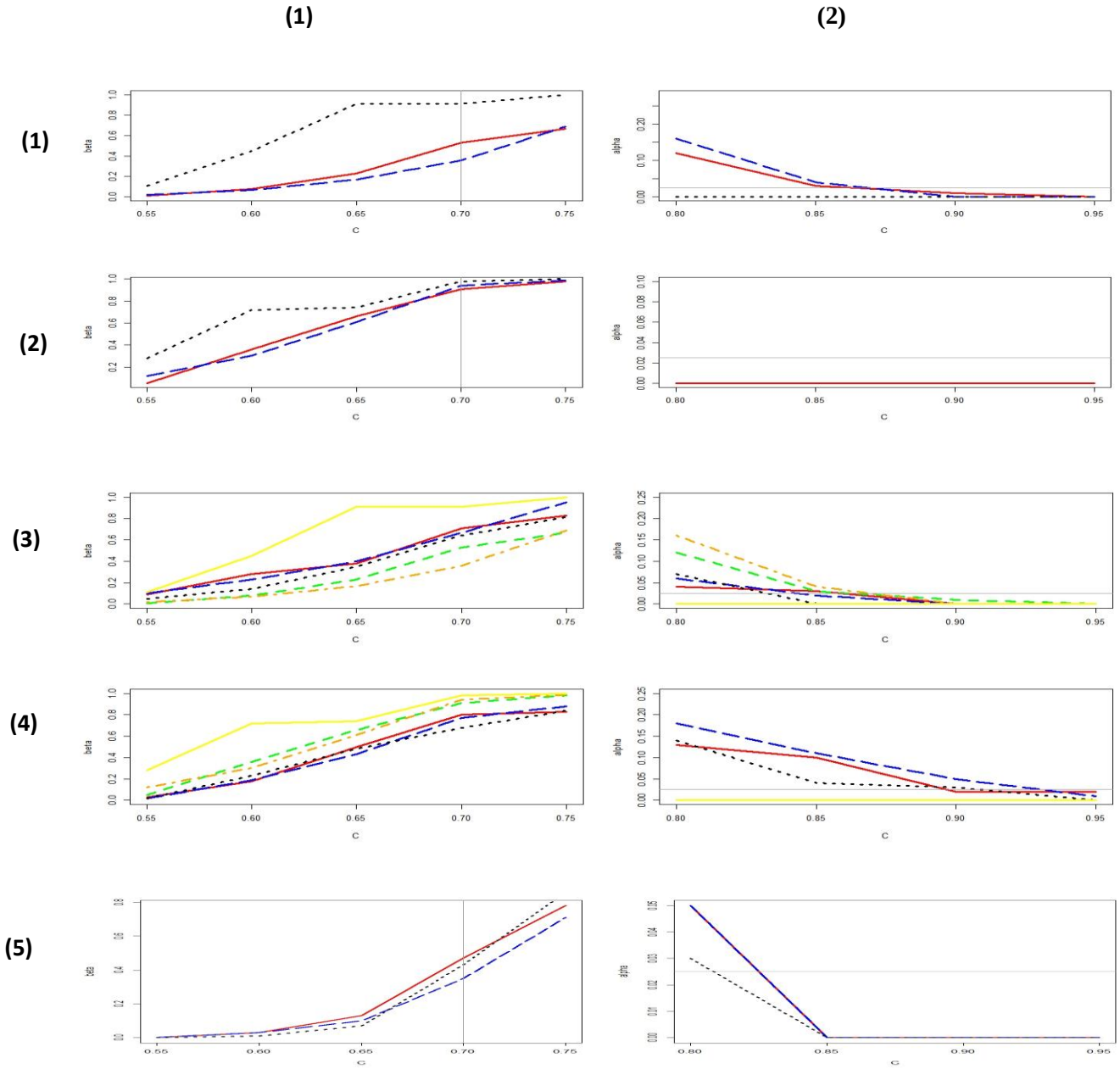


Figure 5: Comparison of "free-floating sharing" versus "hierarchical borrowing" regarding each of the scenarios symbolized by red ("C_{01} = 0.60, C_{02} = 0.80, separate historical trials"), blue ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials"), black ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials, free-floating sharing") for all graphs except (3,·). In graphs (3,1), (3,2), (3,3), (3,4): red ("C_{01} = 0.60, C_{02} = 0.80, separate historical trials, covariate-adjusted synthesis"), blue ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials, covariate-adjusted synthesis"), black ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials, free-floating sharing, covariate-adjusted synthesis"), green ("C_{01} = 0.60, C_{02} = 0.80, separate historical trials, baseline synthesis"), orange ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials, baseline synthesis"), yellow ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials, free-floating sharing, baseline synthesis"). The horizontal line corresponds to a=0.025, and vertical line corresponds to C_0 = 0.7. y-axis = beta/alpha error, x-axis ="C" (active control).

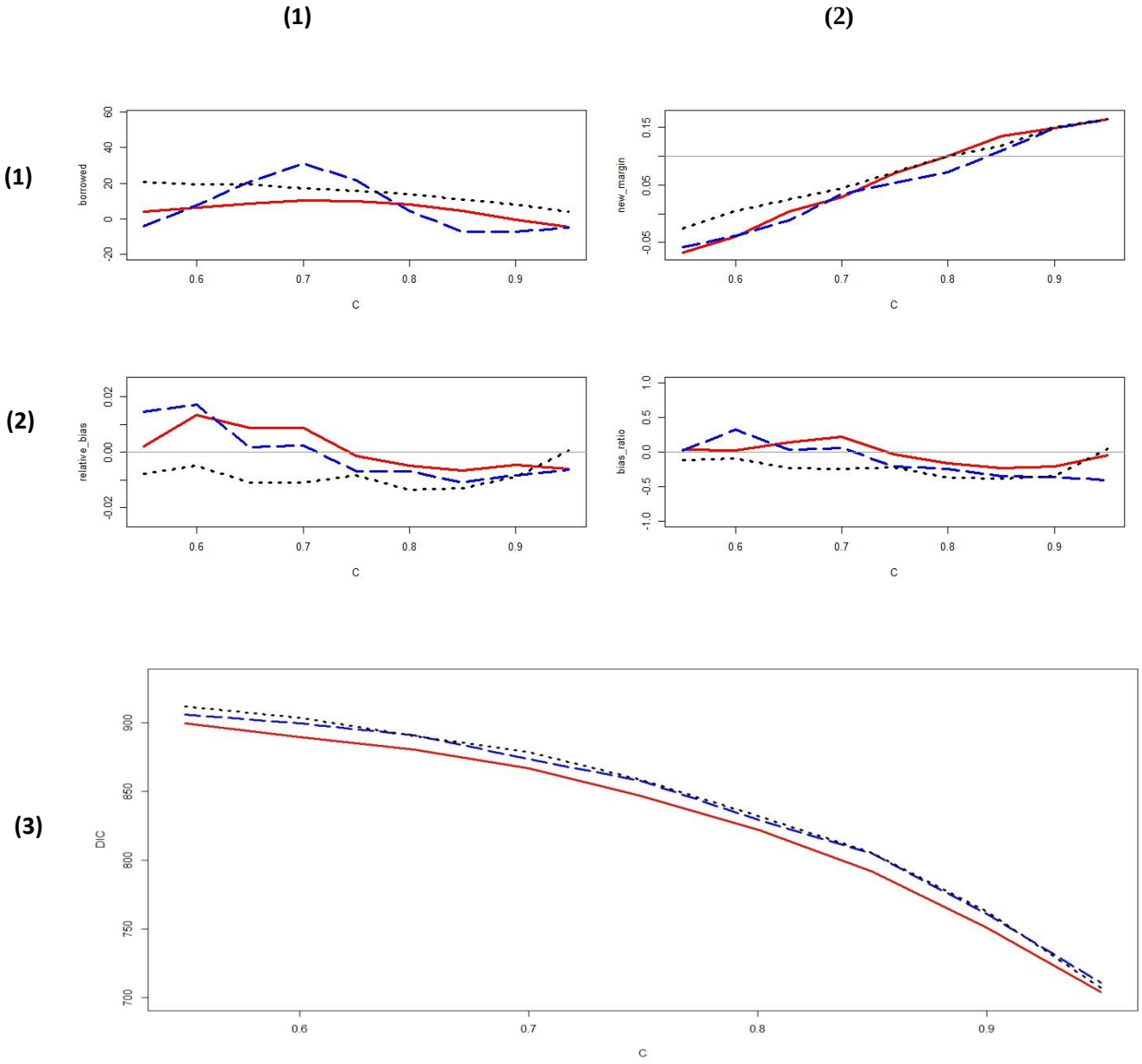


Figure 6: Model diagnostics for comparing "free-floating sharing" versus "hierarchical borrowing," where red ("C_{01} = 0.60, C_{02} = 0.80, separate historical trials"), blue ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials"), black ("C_{01} = 0.60, C_{02} = 0.80, pooled historical trials, free-floating sharing"). (1, 1) = "Number Borrowed", (1,2) = "Adjusted, New Margin after Covariate Adjustment", (2,1) = "Relative Bias", (2,2) = "Bias Ratio", (3,·) = "DIC." Horizontal lines correspond to $Z = 0$ (relative bias and bias ratio), $\delta_{adjusted} = 0.10$ (new margin), vertical line corresponds to $C_0 = 0.7$. y-axis = diagnostic of interest, x-axis ="C" (active control).

6. Conclusions

As C is closer to C_0, borrowing more subjects (e.g., full borrowing) is beneficial. However, borrowing unnecessarily (especially as C drifts away from C_0) increases both Type I and II error in addition to bias. In terms of the deviance information criteria (DIC), the weighted hierarchical performs the best (in terms of lowest DIC) in most cases, except where $C = C_0$ (in this case, full borrowing has the advantage) and at the most extreme points (e.g., C=0.95 >> C_0=0.70), where the unweighted hierarchical method has a slight advantage.

When C is very far away from C_0, the "no borrowing" method performs most optimally in terms of Type I and II error and bias. In fact, borrowing observations between C and C_0 introduces a certain degree of bias which is especially pronounced when there is little constancy between the active and historical control rates.

Overall, if only one borrowing method can be implemented for all values of C, the "weighted hierarchical borrowing" could be considered as the safest choice. It borrows few subjects when $C >> C_0$ or $C << C_0$ but borrows more subjects while accounting for imbalance between sample sizes when $C=C_0$.

When evaluating non-inferiority methods, we can observe that the synthesis method performs better for reducing the amount of Type II (beta) error than the fixed margin methods. However, for Type I (alpha) error, the reverse was true.

The $T - C$ method exhibits less Type II error when compared to the synthesis and fixed margin methods, and shows less Type I error than the synthesis method, but not the fixed margin method. When examining the results, one should take into consideration that the placebo arm is set at $P = 0.50$, which is not very far from the historical control $C_0 = 0.70$. As a result, there will be less $C_0 - P_0$ effect, leading to declare non-inferiority less often.

The covariate adjustment (frequentist-based) for the synthesis method is useful for reducing the Type I error, while the covariate adjustment for the fixed margin method is useful for reducing the Type II error. However, applying the covariate adjustment for the synthesis method for further reducing the Type II error and applying the covariate adjustment for the fixed margin for further reducing the Type I error is not particularly useful.

When comparing separate historical trials versus pooled historical trials, most cases show that the pooled model performs better in terms of beta and alpha error, except at the boundary points where non-inferiority versus inferiority "truth" is not clear (e.g., $C = 0.75 - 0.85$). In terms of hierarchical borrowing, the pooled model borrows more information when C is closer to C_0 compared to the other cases.

For the non-pooled models (separate historical trials), the case of $C_{01} = 0.65$, $C_{02} = 0.75$, which corresponds to moderate difference in the underlying active control rates from trial to trial, contains less Type I and Type II error. However, in terms of selection criteria according to model fit, the non-pooled case of $C_{01} = 0.60$, $C_{02} = 0.80$ performs slightly better (lower DIC) than the others (which are roughly equal). When comparing the non-pooled models to the pooled models, the bias for all model fits is well-controlled.

The final model (free-floating sharing) tends to be more conservative than the hierarchical model. The Type II error is inflated, while the Type I error is consistently well-controlled and is never above the threshold of 0.025. This conclusion applies to both the synthesis and fixed margin methods.

Applying the covariate-adjustment for the synthesis method improves the beta error, although it is still a little inflated compared to the hierarchical model for both the pooled and unpooled analysis. The covariate-adjustment applied to the fixed margin method improves the beta error to such an extent that from $C = 0.67$ to 0.75, this method outperforms the hierarchical model. The alpha error for the fixed margin method (but not the synthesis method) performs well without needing covariate adjustment.

The free-floating sharing modeling approach is the most beneficial when there is little previous information as it controls the alpha error while the inflated beta error resulting from the conservativeness of this modeling approach can be reduced via the covariate-adjustment method under either the fixed margin or synthesis methods. In the case of either inconsistent or nonreliable historical information, the free-floating approach applied to the fixed margin method would be most ideal.

References

1) CBER and CDER FDA Memorandum. Guidance for Industry: Non-Inferiority Clinical Trials, March 2010.

2) CBER and CDER FDA Memorandum. Guidance for Industry: Antibacterial Therapies for Patients with Unmet Medical Need for the Treatment of Serious Bacterial Diseases, July 2013.

3) 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.

4) Dixon and Simon. Bayesian subset analysis. *Biometrics* 1991; **47**: 871-81.

5) Jones, Ohlssen, Neuenschwander, Racine, and Branson. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 2011; **8**:129.

6) Lan KKG, DeMets DL. Design and analysis of group sequential tests based on the type I error spending function. *Biometrika* 1983; **74**: 149-154

7) Nie, L and Soon G. A covariate-adjustment regression model approach to noninferiority margin definition. *Statistics in Medicine* 2010; **29**: 1107-1113.

8) Rothmann M, Wiens B, Chan I. Design and Analysis of Non-Inferiority Clinical Trials, CRC Press, Boca Raton, FL 2012; **5**: 117-120.

9) Simon R. Bayesian design and analysis of active controlled clinical trials. *Biometrics* 1999; **55***:* 484-487.

10) Viele, Berry, Neuenschwander, Amzal, Chen, Enas, Hobbs, Ibrahim, Kinnersley, Lindborg, Micallef, Roychoudhury, Thompson. Use of historical control data for assessing treatment effect in clinical trials. *Pharmaceutical Statistics* 2013; Wiley Online Library DOI: 10.1002 / pst. 1589