Biologics: Targets and Therapy Dovepress

Open Access Full Text Article METHODOLOGY

Comparison of Innovative and Conventional Methods in Biosimilar Bridging Studies with Multiple References

Annpey Pong[1], Susan S Chow[2], Shein-Chung Chow[3]

[1]Biostatistics, Merck & Co Inc., Rahway, NJ, USA; [2]Internal Medicine, The Wright Center for Community Health, Scranton, PA, USA; [3]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

Correspondence: Annpey Pong, Biostatistics, Merck & Co Inc., Rahway, NJ, USA, Email annpey.pong@merck.com

Abstract: For assessment of biosimilar drug products, if there are multiple-reference products (eg, a US-licensed product and an EU-approved product), a biosimilar bridging study with a 3-way pairwise comparison is often conducted. In our paper, two innovative methods in biosimilar bridging study are compared with the conventional method of pairwise comparisons. For parallel study design, the simultaneous confidence interval (CI) method is compared to the convention method. For crossover study design, the multiplicity-adjusted Schuirmann's two one-sided tests (MATOST) is considered. This paper conclude that the simultaneous CI method achieves the similar statistical power to the conventional approach in biosimilarity assessment. However, the MATOST method using the conservative Holm and Bonferroni approaches is not favorable since it leads to a large sample size although it controls the type I error rate.

Keywords: bioequivalence and biosimilarity, multiple references, simultaneous confidence interval approach, multiplicity-adjusted two one-sided tests, MATOST

Introduction

For assessment of biosimilarity between a proposed biosimilar (test) product and an innovative biosimilar (reference) product, FDA and EMA have similar but different approaches. For example, both FDA and EMA require biosimilar products to be highly similar to the reference product in terms of quality characteristics, biological activity, safety and efficacy.[1–5] When there are multiple references (eg, a US-licensed product and an EU-approved reference), the regulatory approval process and specific requirements in the US and EU are different. To be more specific, for a biosimilar product that has both US-licensed and EU-approved reference products, the sponsors are required to conduct bridging studies for biosimilar product vs US-licensed reference product, biosimilar product vs EU-approved reference product, and US-licensed vs EU-approved reference product. Due to the different approval processes and requirements, not all biosimilar products are approved by FDA and EMA using the same clinical trial. Jung et al[6] collected information about 10 biosimilar products and found that half of these biosimilar products have used different clinical trials for approval in the US after being approved in the EU. Additionally, from the bridging studies conducted by Tu et al[7] 3 out of 31 studies (accounting for two biosimilar products) failed to prove the similarity between US-licensed and EU-approved reference products.

In addition, both FDA and EMA have different requirements for using foreign approved comparators (FACs). If the sponsor wants to use FAC instead of domestic reference biologic product (RBP), intuitively, they need to prove that FAC is materially representative of the domestic RBP.[8] EMA permits FAC in regulator submission, but highly recommends that the FAC be licensed by a country adopting guidelines from the International Conference on Harmonization (ICH), whereas the FDA always requires a 3-way pairwise comparative bridging study.[9]

Received: 17 July 2024
Accepted: 29 November 2024
Published: 6 December 2024

The conventional 3-way pairwise comparison has the following issues. (i) The determination of equivalence margins without taking the variability into consideration, (ii) Not using all collected data in each pairwise comparison, and (iii) The issue of multiple testing. To overcome these issues, several methods have been proposed in the literature. For example, Zheng et al[10] proposed a simultaneous confidence interval (CI) approach based on fiducial inference in parallel group biosimilar bridging studies with multiple references. To alleviate type I error inflation in multiple testing, Zheng et al[11] proposed using a multiplicity-adjusted TOST (MATOST) which utilizes some p-value adjustment methods.

This article aims to further investigate the favorable properties of the two innovative approaches, ie, the simultaneous CI method and MATOST compared with conventional pairwise comparison. Specifically, we study the statistical powers of using the simultaneous CI method, MATOST and the conventional pairwise comparison in parallel design and crossover design. This paper can provide more information to researchers on bioequivalence tests with multiple references. In the next section, more detailed information about the regulatory requirement for biosimilar products is provided. Methods for Multiple Comparisons discusses the disadvantages of conventional pairwise comparison and introduces the simultaneous CI method in parallel design and the MATOST in crossover design, specifically, the Williams design for comparing more than two treatments. Simulation Study contains the results from the simulation that is conducted to evaluate the performance among methods. Finally, Results provides the concluding remarks.

Regulatory Requirement

A biosimilar product is often defined as a biological product that is highly similar to the reference product notwithstanding the minor difference in clinically inactive components, ie, there is no clinically meaningful difference between the biosimilar and reference products in terms of safety, purity and potency.[9] FDA requires biosimilar products to have no clinically meaningful differences compared with the reference product in terms of both effectiveness and safety; and EMA defines biosimilar product as a biological medicinal product containing the same active substance as the reference product, and similar in terms of "quality characteristics, biological activity, safety and efficacy". For biosimilar product approval, FDA proposed a guideline on a stepwise approach for obtaining totality-of-The-evidence, ie, a step-by-step approach to collect global biosimilarity evidence across different domains.[3,12] This stepwise approach can be summarized into three key components: (i) analytical similarity assessment, investigating the structural/functional characteristic of the biosimilar product via critical quality attributes (CQAs) at various stages of the manufacturing process; (ii) pharmacokinetics (PK) and pharmacodynamics (PD) similarity assessment, demonstrating PK and PD by human clinical pharmacology studies after passing animal toxicity studies; and (iii) clinical similarity assessment, gathering information on immunogenicity, clinical efficacy and safety.

In addition to the stepwise approach, FDA has a rigorous requirement on multiple references. When there are multiple references, for example, the US-licensed reference product and EU-approved reference product, the 3-way comparison is recommended. The main purpose of this requirement is to assure that the quality of the drug substance or product meets product specification, to demonstrate biosimilarity between test product and US-licensed product, and to establish a bridge to justify the use of clinical data generated using EU-approved reference product as the comparator. Since biosimilar products are often large-molecule products made of living cells and organisms, they are highly sensitive to environmental conditions. These strict requirements on biosimilar products are necessary to ensure the efficacy and safety of the biosimilar products appearing on market.

In the guideline on biosimilar product bioequivalence tests for a certain CQA, FDA suggests using $1.5\sigma_R$ as the equivalence test margin (σ_R is the standard deviation of the reference product), ie, the equivalence test hypotheses are[4]

$$H_0 : |\mu_T - \mu_R| \geq \delta \, vs \, H_1 : |\mu_T - \mu_R| < \delta, \tag{1}$$

where $\mu_T - \mu_R$ is the mean difference between the test and the reference products, and $\delta = 1.5\sigma_R$ is the bioequivalence margin. In addition, the estimand of interest can also be the population mean ratio, then the equivalence test hypotheses are

$$H_0 : \frac{\mu_T}{\mu_R} \notin [80\%, 125\%] \, vs \, H_1 : \frac{\mu_T}{\mu_R} \in [80\%, 125\%], \tag{2}$$

where the estimand can also be $\log\frac{\mu_T}{\mu_R}$. The hypothesis testing in Eq. (1) and (2) are the most commonly used in bioequivalent research, to population mean difference as the estimand, researchers must use the domain knowledge or previous literature to determine the boundary.[7,13,14]

This fixed margin approach using σ_R has some disadvantages, especially for multiple-reference scenario. σ_R is always unknown in practice, and simply replacing σ_R by estimated σ_R is unfavorable because it does not take the variability of estimating σ_R using samples into consideration. As for the situation for multiple references, in a 3-way pairwise comparison, the equivalence margin can be incomparable. For example, when comparing biosimilar product vs US-licensed reference and biosimilar product vs EU-licensed reference, the standard deviation of reference product may be different. In this way, the equivalence tests will have different precision.

Methods for Multiple Comparisons

Conventional Method of Pairwise Comparisons

While using a non-US-licensed product in regulatory submission for a biosimilar product, 3-way bridging evidence should be provided by the sponsor to justify the capability of these comparative data from non-US-licensed product in assessing biosimilarity and bridging to the US-licensed reference product. In other words, the pairwise bioequivalence among the biosimilar product, US-licensed reference and non-US-licensed reference (eg, EU-approved reference). The use of conventional pairwise comparison be impaired the following disadvantages. (i) Each pairwise comparison does not fully utilize data collected from the entire study (eg, when comparing the EU-approved reference and biosimilar product, data collected from US-licensed reference are not used). In this way, the results may be biased and hence misleading. (ii) Pairwise comparisons do not use the same reference product in each comparison, and each comparison may use a different equivalence acceptance criterion (EAC) margin. Then, the precisions of each equivalence test are different. (iii) Since there are multiple comparisons, type I error will be inflated unless some significance-level adjustment approaches are considered. (iv) These pairwise comparisons cannot distinguish the following relationship among biosimilar product (T), US-licensed reference (R_1) and EU-approved reference (R_2):

(a) $R_1 > T > R_2$,
(b) $R_1 > R_2 > T$,
(c) $T > R_1 > R_2$,
(d) $T > R_2 > R_1$,
(e) $R_2 > T > R_1$,
(f) $R_2 > R_1 > T$.

To overcome deficiencies (i) and (ii), Zheng et al[10] proposed the simultaneous CI method based on fiducial inference in parallel group design. To overcome deficiency (iii), MATOST is proposed by Zheng et al.[11] More detailed information about these two innovative approaches in biosimilar study with multiple references are provided in this research. Without loss of generality, US-licensed reference product and EU-approved reference product are considered as the two reference products, and US-licensed reference is the primary reference.

Innovative Methods of Bioequivalence Tests

Simultaneous Confidence Interval Method

Consider a three-arm parallel randomized clinical trial with an allocation ratio 1:1:1. Let μ_T, μ_{R_1} and μ_{R_2} denote the population mean for biosimilar product, US-licensed reference and EU-approved reference. The hypotheses of the 3-way pairwise equivalence test are

$$H_{01}: \left|\mu_T - \mu_{R_1}\right| \geq \delta_1 \ vs \ H_{11}: \left|\mu_T - \mu_{R_1}\right| < \delta_1,$$

$$H_{02}: \left|\mu_T - \mu_{R_2}\right| \geq \delta_2 \ vs \ H_{12}: \left|\mu_T - \mu_{R_2}\right| < \delta_2, \quad (3)$$

$$H_{03}: \left|\mu_{R_2} - \mu_{R_1}\right| \geq \delta_3 \ vs \ H_{13}: \left|\mu_{R_2} - \mu_{R_1}\right| < \delta_3, \quad (4)$$

where $\delta_i (i = 1, 2, 3)$ are the equivalence margins. It should be noted that in hypotheses (2) and (3), the reference products are different, which makes the equivalence margin to be incomparable. As mentioned in Regulatory Requirement, using the fixed equivalence margin $(1.5\sigma_R)$ is problematic. Zheng et al[10] proposed the simultaneous CI method based on fiducial inference to overcome this issue.

Assume that each arm independently follows a normal distribution: $N(\mu_T, \sigma_T^2)$, $N\left(\mu_{R_1}, \sigma_{R_1}^2\right)$ and $N\left(\mu_{R_2}, \sigma_{R_2}^2\right)$. Let $\bar{X}_T$, $\bar{X}_{R_1}$ and $\bar{X}_{R_2}$ denote the sample means for each arm, the sample sizes for each arm are n_T, n_{R_1} and n_{R_2}. Note that the data used to estimate standard deviation and conduct bioequivalence tests are different. In other words, samples used to obtain equivalence margins $(1.5\sigma_R)$ will be collected separately from those used to conduct hypothesis testing. Let $\hat{\sigma}_{R_1}$ and $\hat{\sigma}_{R_2}$ denoted the empirical standard deviation, which are computing using n_1 and n_2 samples, respectively. Start with the simplest case for a 3-way comparison, ie, the standard deviations of all three arms are the same.

Case 1. Assume $\sigma_T = \sigma_{R_1} = \sigma_{R_2}$

If the standard deviations are the same, the fiducial probability of all pairwise equivalence hold is

$$FP_{10}(\sigma_{R_1}) = \int\int\int_{|x-y|\le 1.5\sigma_{R_1},\,|x-z|\le 1.5\sigma_{R_1},\,|y-z|\le 1.5\sigma_{R_1}} f(x,y,z)\,dx\,dy\,dz, \tag{5}$$

where $f(x,y,z) = f_1(x)f_2(y)f_3(z)$ is the joint fiducial probability density of $(\mu_T, \mu_{R_1}, \mu_{R_2})$, and

$$\mu_T \sim N\left(\bar{X}_T, \frac{\sigma_{R_1}^2}{n_T}\right), \mu_{R_1} \sim N\left(\bar{X}_{R_1}, \frac{\sigma_{R_1}^2}{n_{R_1}}\right) \text{and } \mu_{R_2} \sim N\left(\bar{X}_{R_2}, \frac{\sigma_{R_1}^2}{n_{R_2}}\right).$$

Let α denote the significance level. We reject H_{01} in Eq. (2) and conclude the bioequivalence between a biosimilar product and a primary biosimilar product (R_1), if $FP_1(\sigma_{R_1}) \ge 1 - \alpha$; otherwise, fail to reject H_{01}. To take the variability of the estimated standard deviation into consideration, the proposed the least favorable version of fiducial probability

$$FP_1\left(\hat{\sigma}_{R_1}^{'}\right) = \int\int\int_{|x-y|\le 1.5\hat{\sigma}_{R_1}^{'},\,|x-z|\le 1.5\hat{\sigma}_{R_1}^{'},\,|y-z|\le 1.5\hat{\sigma}_{R_1}^{'}} f(x,y,z)\,dx\,dy\,dz, \tag{6}$$

where $\hat{\sigma}_{R_1}^{'} = \sqrt{\frac{(n_1-1)\hat{\sigma}_{R_1}^2}{\chi^2_{1-\frac{\alpha}{2}}(n_1-1)}}$.

Case 2. Without assumption of $\sigma_T = \sigma_{R_1} = \sigma_{R_2}$

Treating US-licensed reference product as the primary one, to make the equivalence margins comparable, we can choose to only use one reference product while constructing margins. Then the fiducial probability can be derived as[10]

$$FP_2(\sigma_{R_1}) = \int\int\int_{|x-y|\le 1.5\sigma_{R_1},\,|x-z|\le 1.5\sigma_{R_1},\,|y-z|\le 1.5\sigma_{R_1}} g(x,y,z)\,dx\,dy\,dz, \tag{7}$$

where $g(x,y,z) = f_1(x)g_2(y)g_3(z)$ is the joint fiducial probability density of $\left(\mu_T, \mu_{R_1}, \mu_{R_2}\right)$, and

$$\mu_{R_1} \sim \bar{X}_{R_1} + \frac{\tilde{\sigma}_{R_1} t\left(n_{R_1} - 1\right)}{\sqrt{n_{R_1}}},$$

$$\mu_{R_2} \sim \bar{X}_{R_2} + \frac{\tilde{\sigma}_{R_2} t\left(n_{R_2} - 1\right)}{\sqrt{n_{R_2}}},$$

where $\tilde{\sigma}_{R1} = \left(\sum_{i=1}^{n_{R1}} \frac{\left(X_{iR1} - \bar{X}_{R1}\right)^2}{n_{R1}-1}\right)^{1/2}$ and $\tilde{\sigma}_{R_2} = \left(\sum_{i=1}^{n_{R_2}} \frac{\left(X_{iR_2} - \bar{X}_{R_2}\right)^2}{n_{R_2}-1}\right)^{1/2}$.

Let $\hat{\sigma}_1 = \frac{\tilde{\sigma}_{R_1}}{\sqrt{n_{R_1}}}$ and $\hat{\sigma}_2 = \frac{\tilde{\sigma}_{R_2}}{\sqrt{n_{R_2}}}$, then

$$g_2(y) = \frac{1}{\hat{\sigma}_1} f_{t\left(n_{R_1}-1\right)}\left(\frac{y - \bar{X}_{R_1}}{\hat{\sigma}_1}\right) \text{and } g_3(z) = \frac{1}{\hat{\sigma}_2} f_{t\left(n_{R_2}-1\right)}\left(\frac{z - \bar{X}_{R_2}}{\hat{\sigma}_2}\right), \tag{8}$$

where $f_{t\left(n_{R_1}-1\right)}(\cdot)$ and $f_{t\left(n_{R_2}-1\right)}(\cdot)$ are the density function of t distribution with degree of freedom $n_{R_1} - 1$ and $n_{R_2} - 1$, respectively. The least favorable version of fiducial probability is

$$FP_2\left(\hat{\sigma}'_{R_1}\right) = \int\int\int_{|x-y|\le 1.5\sigma'_{R_1},|x-z|\le 1.5\sigma'_{R_1},|y-z|\le 1.5\sigma'_{R_1}} g(x,y,z)\,dx\,dy\,dz. \tag{9}$$

On the other hand, we may choose to use two reference products for the equivalence margin; however, this approach cannot make equivalence margins comparable. Then, the fiducial probability can be written as

$$FP_3(\sigma_{R_1},\sigma_{R_2}) = \int\int\int_{|x-y|\le 1.5\sigma_{R_1},|x-z|\le 1.5\sigma_{R_2},|y-z|\le 1.5\sigma_{R_1}} h(x,y,z)dxdydz, \tag{10}$$

where $h(x,y,z) = f_1(x)f_2(y)g_3(z)$. And the least favorable version of fiducial probability is

$$FP_3\left(\hat{\sigma}''_{R_1},\ \hat{\sigma}''_{R_2}\right) = \int\int\int_{|x-y|\le 1.5\hat{\sigma}''_{R_1},|x-z|\le 1.5\hat{\sigma}''_{R_2},|y-z|\le 1.5\hat{\sigma}''_{R_1}} h(x,y,z)\,dx\,dy\,dz, \tag{11}$$

where $\hat{\sigma}''_{R_1} = \sqrt{\frac{(n_1-1)\hat{\sigma}_{R_1}}{\chi^2_{1-\frac{1-\sqrt{1-\alpha}}{2}}(n_1-1)}}$ and $\hat{\sigma}''_{R_2} = \sqrt{\frac{(n_2-1)\hat{\sigma}_{R_2}}{\chi^2_{1-\frac{1-\sqrt{1-\alpha}}{2}}(n_2-1)}}$.

Multiplicity-Adjusted Schuirmann's TOST Approach

The parallel group design has the drawback of being unable to control the inter-subject variability, which weakens the interchangeability evaluation of biosimilar products. To address this issue, crossover designs may be considered. The Williams design (a special type of crossover design) can successfully reduce the inter-subject variability and maintain variance-balanced, which may ensure the estimated pairwise formulation effects with the same precision. The treatment assignment of one Williams design considered in this section is shown in Table 1.

The crossover design model and basic assumptions are

$$Y_{ijk} = \log(Z_{ijk}) = \mu + G_k + S_{ik} + P_j + F_{(j,k)} + C_{(j-1,k)} + e_{ijk}, \tag{12}$$

where Y_{ijk} is the response for every patient;

μ is the overall mean;

G_k is the fixed effect of kth sequence, $k = 1,\ldots,K$, and $K = 6$;

P_j is the fixed effect of jth period, $j = 1,\ldots,J$, and $J = 3$;

$F_{(j,k)}$ is the fixed formulation effect of k^{th} sequence at j^{th} period $\left(\sum_{j,k}F_{(j,k)} = 0\right)$;

$C_{(j-1,k)}$ is the fixed first-order carry over effect $\left(\sum_{j,k}C_{(j-1,k)} = 0\right)$;

S_{ik} is the random effects of i^{th} subject from k^{th} sequence $\left(S_{ik}\sim\left(0,\sigma_s^2\right)\right)$, e_{ijk} is the random error $\left(e_{ijk}\sim\left(0,\sigma_e^2\right)\right)$, and $i = 1,\ldots,n_k$.

In the bioequivalence test, the primary parameter of interest is the treatment effect. The population treatment effect for each arm, ie μ_T, μ_{R_1} and μ_{R_2}, can be estimated as following:

$$\hat{\mu}_T = \hat{\mu} + \hat{F}_T, \hat{\mu}_{R_1} = \hat{\mu} + \hat{F}_{R_1}\ and\ \hat{\mu}_{R_2} = \hat{\mu} + \hat{F}_{R_2},$$

Table 1 Treatment Assignment in Williams Design

Sequence	Period		
	1	2	3
1	T	R_2	R_1
2	R_1	T	R_2
3	R_2	R_1	T
4	R_1	R_2	T
5	R_2	T	R_1
6	T	R_1	R_2

where $\hat{\mu} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{jk}$ and $\hat{F}_h$ can be determined by ordinary least square (OLS).

Let the parameters of interest in OLS denoted by $\beta = [\mu P_1 P_2 G_1 G_2 G_3 G_4 G_5 F_T F_{R_1} C_T C_{R_1}]'$. The OLS estimator of β is

$$\hat{\beta}_{LS} = \boldsymbol{A}\bar{\boldsymbol{y}}, \tag{13}$$

where $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{X}$ is the design matrix (as shown in Table 2), $\bar{\boldsymbol{y}} = [\bar{\boldsymbol{y}}_1\ \bar{\boldsymbol{y}}_2\ \ldots\ \bar{\boldsymbol{y}}_{\boldsymbol{K}}]'$, $\bar{\boldsymbol{y}}_{\boldsymbol{k}} = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{y}_{\boldsymbol{ik}}$ and $\boldsymbol{y}_{\boldsymbol{ik}} = [y_{i1k}\ y_{i2k}\ y_{i3k}]'$. Since the unbiased estimators of every formulation effect can be written as a linear combination of $\bar{y}_{jk}$, so does the unbiased estimator for formulation effect contrasts. The linear combination has the format:

$$L_a = \sum_{k=1}^{K}\sum_{j=1}^{J} C_{ajk}\bar{y}_{jk}, \tag{14}$$

where C_{ajk} is the linear combination coefficient. Under normality assumption, ie, $\sum_j C_{ajk} = \sum_k C_{ajk} = 0$, the variance of L is

$$var(L_a) = \sigma_e^2\sum_{k=1}^{K}\frac{1}{n_k}\sum_{j=1}^{J} C_{ajk}^2. \tag{15}$$

The coefficients for estimated formulations and estimated formulation effect contrasts can be found in Tables 3 and 4, respectively.

In addition to using treatment effect difference as the estimand, the treatment effect ratio can also be a valid candidate for the estimand. The equivalence hypothesis testing's hypotheses can also be written as

$$H_{01} : \log\left(\frac{\mu_T}{\mu_{R_1}}\right) \notin (\log(\theta_L), \log(\theta_U))\ vs\ H_{11} : \log\left(\frac{\mu_T}{\mu_{R_1}}\right) \in (\log(\theta_L), \log(\theta_U)), \tag{16}$$

$$H_{02} : \log\left(\frac{\mu_T}{\mu_{R_2}}\right) \notin (\log(\theta_L), \log(\theta_U))\ vs\ H_{12} : \log\left(\frac{\mu_T}{\mu_{R_2}}\right) \in (\log(\theta_L), \log(\theta_U)), \tag{17}$$

$$H_{03} : \log\left(\frac{\mu_{R_2}}{\mu_{R_1}}\right) \notin (\log(\theta_L), \log(\theta_U))\ vs\ H_{13} : \log\left(\frac{\mu_{R_2}}{\mu_{R_1}}\right) \in (\log(\theta_L), \log(\theta_U)), \tag{18}$$

Table 2 The Design Matrix ($\boldsymbol{X}$) for Williams Design

(j, k)	μ	P_1	P_2	G_1	G_2	G_3	G_4	G_5	F_T	F_{R1}	C_T	C_{R1}
(1,1)	1	1	0	1	0	0	0	0	1	0	0	0
(2,1)	1	0	1	1	0	0	0	0	−1	−1	1	0
(3,1)	1	−1	−1	1	0	0	0	0	0	1	−1	−1
(1,2)	1	1	0	0	1	0	0	0	0	1	0	0
(2,2)	1	0	1	0	1	0	0	0	1	0	0	1
(3,2)	1	−1	−1	0	1	0	0	0	−1	−1	1	0
(1,3)	1	1	0	0	0	1	0	0	−1	−1	0	0
(2,3)	1	0	1	0	0	1	0	0	0	1	−1	−1
(3,3)	1	−1	−1	0	0	1	0	0	1	0	0	1
(1,4)	1	1	0	0	0	0	1	0	0	1	0	0
(2,4)	1	0	1	0	0	0	1	0	−1	−1	0	1
(3,4)	1	−1	−1	0	0	0	1	0	1	0	−1	−1
(1,5)	1	1	0	0	0	0	0	1	−1	−1	0	0
(2,5)	1	0	1	0	0	0	0	1	1	0	−1	−1
(3,5)	1	−1	−1	0	0	0	0	1	0	1	1	0
(1,6)	1	1	0	−1	−1	−1	−1	−1	1	0	0	0
(2,6)	1	0	1	−1	−1	−1	−1	−1	0	1	1	0
(3,6)	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	1

Table 3 Coefficients (C_{ajk}) to Estimate Formulation Effects

Sequence	F_T			F_{R1}			F_{R2}		
	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
1	3	0	−3	−1	−2	3	−2	2	0
2	−2	2	0	3	0	−3	−1	−2	3
3	−1	−2	3	−2	2	0	3	0	−3
4	−1	−2	3	3	0	−3	−2	2	0
5	−2	2	0	−1	−2	3	3	0	−3
6	3	0	−3	−2	2	0	−1	−2	3

Notes: Coefficients are multiplied by 24. P_j (j = 1,2,3) represents the fixed effect for jth period. F_h ($h \in \{T, R_1, R_2\}$) represents the fixed effect for formulation h.

Table 4 Coefficients (C_{ajk}) to Estimate Formulation Effect Differences

Sequence	$\theta_{TR1} = F_T - F_{R1}$			$\theta_{TR2} = F_T - F_{R2}$			$\theta_{R2R1} = F_{R2} - F_{R1}$		
	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
1	4	2	−6	5	−2	−3	−1	4	−3
2	−5	2	3	−1	4	−3	−4	−2	6
3	1	−4	3	−4	−2	6	5	−2	−3
4	−4	−2	6	1	−4	3	−5	2	3
5	−1	4	−3	−5	2	3	4	2	−6
6	5	−2	−3	4	2	−6	1	−4	3

Notes: Coefficients are multiplied by 24. Notations are the same as Table 3.

where θ_L and θ_U are the equivalence margin, usually set as 80% and 125%. The test statistics are

$$T_L = \frac{\log\left(\frac{\hat{\mu}_h}{\hat{\mu}_{h'}}\right) - \log(\theta_L)}{\left[S^2 \sum_{k=1}^{K} \frac{1}{n_k} \sum_{j=1}^{J} C_{ajk}^2\right]^{\frac{1}{2}}} \text{ and } T_U = \frac{\log\left(\frac{\hat{\mu}_h}{\hat{\mu}_{h'}}\right) - \log(\theta_U)}{\left[S^2 \sum_{k=1}^{K} \frac{1}{n_k} \sum_{j=1}^{J} C_{ajk}^2\right]^{\frac{1}{2}}}, \tag{19}$$

where S^2 is the intra-subject mean square error (MSE), C_{ajk}^2 are the contrast coefficients in Eq. (15), and $h, h' \in \{T, R_1, R_2\} (h \neq h')$. Under null hypothesis, T_L and T_U follows t distribution with degree of freedom $v = 2(N - 3)$, where $N = \sum_{k=1}^{K} n_k$. Claim equivalence between T, R_1 and R_2, if all test statistics satisfy

$$T_L > t_\alpha(v) \text{ and } T_U < t_{1-\alpha}(v), \tag{20}$$

where $t_\alpha(v)$ is the $\alpha\%$ percentile of t distribution with degree of freedom v. The p-value of TOST can be computed as following (Lakens, 2017)

$$p = \max\{1 - F(T_L), F(T_U)\}, \tag{21}$$

where $F(\cdot)$ is the cumulative distribution function (CDF) of t distribution with degree of freedom v. However, in the 3-way pairwise comparison using TOST, multiple-testing problem may affect the overall type I error rate. Zheng et al[11] proposed the multiplicity-adjusted TOST (MATOST) to alleviate type I error inflation. Thus, the p-value computed in Eq. (21) needs to be adjusted using an adjustment approach, such as Holm adjustment and Bonferroni's adjustment.

Simulation Study

Simulation studies are conducted aiming to investigate the performance of simultaneous CI methods in parallel design while using mean difference as the estimand and the performance of TOST and MATOST in Williams design while using

the population mean ratio as the estimand. The significance level is $\alpha = 0.1$. The equivalence margins are either $1.5\sigma_R$ or 80–125%. Each simulation contains 1000 iterations.

To investigate the performance of the simultaneous CI method proposed by Zheng et al[10] in parallel group design for 3-way pairwise comparison, a simulation study is conducted comparing statistical power in equivalence tests. Specifically, 3 types of simultaneous CI methods (denoted by FP1, FP2 and FP3) are considered: (i) FP1: assume $\sigma_T = \sigma_{R_1} = \sigma_{R_2}$, (ii) FP2: without equal standard deviation assumption using σ_{R_1} in equivalence margin, and (iii) FP3: without equal standard deviation assumption using σ_{R_1} and σ_{R_2} in equivalence margin. The conventional pairwise comparison is conducted using ANOVA (Analysis of Variance). The simulations are performed on a range of sample size from 10 to 100 per arm.

As shown in Figure 1, among all four assumptions from (A) to (D), FP1 achieves the highest statistical power when the sample size per arm is small. Even when the assumption about equal variability is inaccurate, FP1 can still achieve the desired power. When the true standard deviations of all three arms are the same, the conventional method has the poorest performance; however, when the standard deviations are different, FP1's performance is better than FP2 and FP3. When the sample size is larger than 20, the statistical powers of all methods are greater than the desired 80%. In general, while taking the variability of estimation into consideration, the simultaneous CI method can achieve similar power as the conventional approach.

In the simulation study for TOST in Williams design, the statistical powers under different sample sizes with and without p-value adjustment are compared. The p-value adjustment method considered here is the MATOST approach proposed by Zheng et al[12], specifically, the Bonferroni method and the Holm method. The simulations are performed on a range of sample size from 10 to 100 per sequence and 4 coefficient of variations $CV \in \{0.1, 0.2, 0.3, 0.4\}$ with 3 assumptions on $\{\mu_T, \mu_{R1}, \mu_{R2}\}$.

As shown in Figure 2, the Bonferroni's method and the Holm method are almost identical in terms of TOST's statistical power. The statistical power of tests without using any type I error adjustment method is higher compared with the one with adjustments. When the coefficient of variation increases, the statistical power difference between TOST with and without using p-value adjustment becomes smaller. And the required sample size to maintain enough power increases when CV increases, though the impact of CV on the sample size per arm is not great.

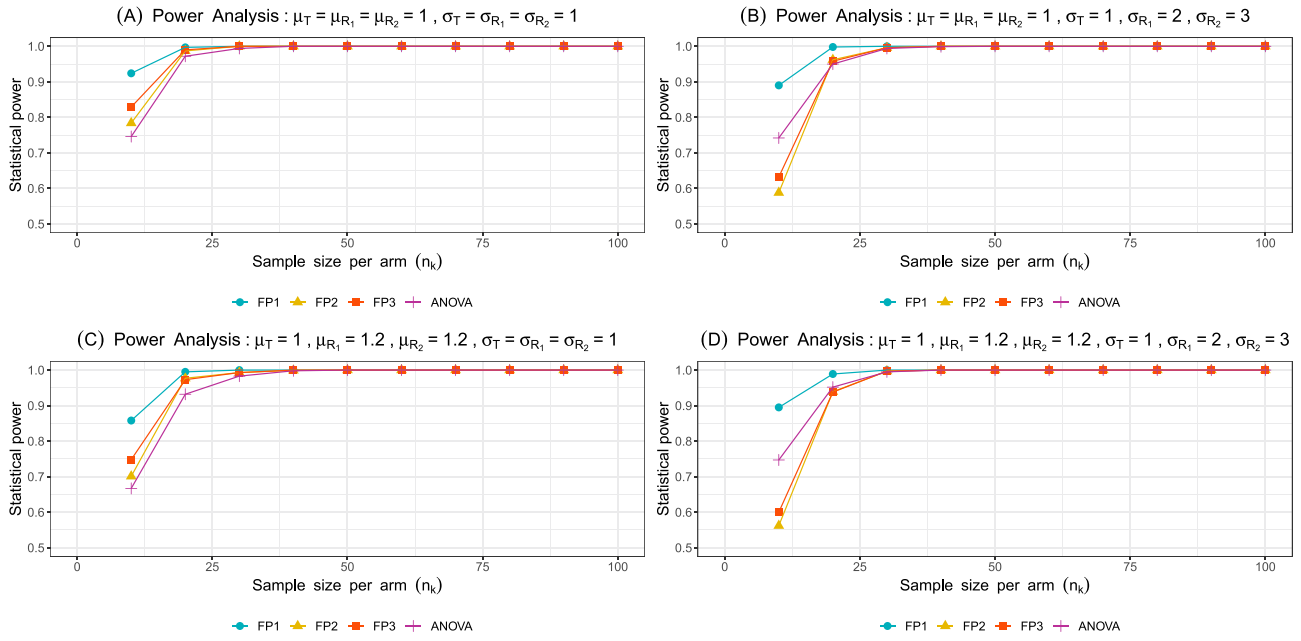

Figure 1 Power analysis for comparing simultaneous CI method versus 3-way pairwise comparison in parallel design. Sub-figures (**A–D**) represent the 4 different assumptions on (μ_T, μ_{R1}, μ_{R2}) and (σ_T, σ_{R1}, σ_{R2}).

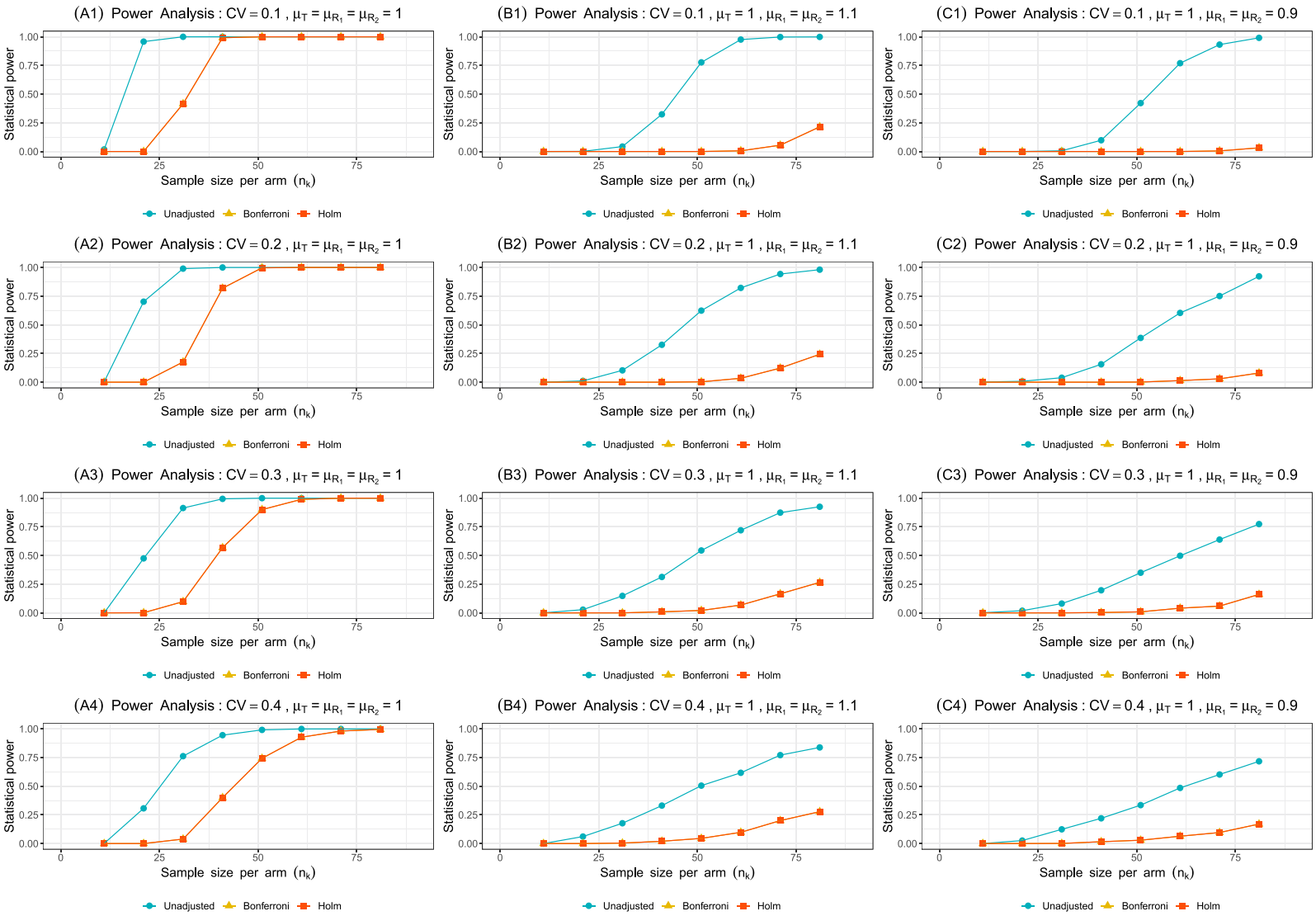

Figure 2 Power analysis for comparing Schuirmann's TOST method versus 3-way pairwise comparison in crossover design. Sub-figures (**A1**–**A4**) represents CV from 0.1 to 0.4 when μ_T = μ_{R1} = μ_{R2} = 1; (**B1**–**B4**) represents CV from 0.1 to 0.4 when μ_T = μ_{R1} = μ_{R2} = 1.1; (**C1**–**C4**) represents CV from 0.1 to 0.4 when μ_T = μ_{R1} = μ_{R2} = 0.9.

https://doi.org/10.2147/BTT.S470182

When the true population means for each arm are the same, when $CV \in \{0.1, 0.2, 0.3\}$, setting $n_k = 50$ is enough to maintain 80% power. However, when $\frac{\mu_T}{\mu_{R_1}} = \frac{\mu_T}{\mu_{R_2}} = \frac{1}{1.1}$, let the sample size per arm be 75, the power is less than 0.25; when $\frac{\mu_T}{\mu_{R_1}} = \frac{\mu_T}{\mu_{R_2}} = \frac{1}{0.9}$, let the sample size per arm be 75, the power is less than 0.20. In other words, the MATOST is too conservative, ie, more likely to accept the null hypothesis, for setting B and C in Figure 2. Thus, the price for controlling type I error rate is to increase the sample size to a very high level while maintaining enough statistical power. Since there are 6 sequences, the overall sample size will be difficult to accrue in practice.

Results

To compare the innovative methods with the conventional method in biosimilar bridging studies with multiple references, the results are summarized as below.

In a parallel study design, the simultaneous CI method based on Zheng et al[10] using fiducial probabilities was performed to conduct pairwise equivalence tests. In the simulation study, four methods are considered for pairwise comparison (FP1, FP2, FP3, and ANOVA) which include three simultaneous CI methods with different assumptions and one conventional approach. The results showed that all approaches can maintain desired power when the sample size per arm is larger than 20. Among these methods, the simultaneous CI method under equal variability has the best performance. All simultaneous CI methods can achieve similar statistical power to the conventional approach. This concludes that the simultaneous CI method works well in parallel study design.

In crossover study, Williams design that often used in bioequivalence trial was applied. Three approaches were considered for power calculation for MATOST method. They are the Holm and Bonferroni approaches with type 1 error adjustment, and the 3rd approach without p-value adjustment.

From simulation results, the Holm method and Bonferroni method are very conservative, ie, TOST is more likely to fail to reject the null hypothesis to control type I error inflation. When the coefficient of variation is large, the sample size for each arm requires an undesirable high level (more than 50 per arm to achieve 75% power). Thus, in a crossover design with multiple formulations, especially the Williams design, the MATOST method is not favorable since it may lead to an unrealistic large sample size, though it can control the type I error rate.

Discussions

For biosimilar product regulatory submission, FDA requires sponsors to conduct a 3-way pairwise comparison if both US-licensed reference and EU-approved reference exists. In equivalence tests, determining the equivalence margin is of great importance. Although FDA's guideline recommends using $1.5\sigma_R$ as the equivalence margin, σ_R is unknown in practice. Using sample estimation in the hypotheses setting is inappropriate since it does not take the variability of estimation into consideration.

In this article, although the Williams design is the primary focus, the MATOST can also be applied to the complete n-of-1 design. For future research, it is possible to consider further utilization of fiducial inference under crossover designs and generalize the simultaneous CI method proposed by Zheng et al.[10] Generalizing the simultaneous CI method can help researchers in determining equivalence margins by sampling standard deviations and taking the variability of estimation into consideration. Additionally, another type I error inflation controlling method needs to be proposed specifically for crossover design to reduce large sample size for conducting a biosimilar study.

Acknowledgment

The authors would like to thank Ms. Peijin Wang for her considerable assistance in the simulation study.

Disclosure

Pong A is the employees of Merck & Co., Inc., Rahway, NJ, USA, who own stock in Merck & Co., Inc.. The authors report no conflicts of interest in this work.

References

1. EMA. European Medicines Agency. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-similar-biological-medicinal-products-rev1_en.pdf. Accessed December 27, 2022.
2. EMA. *Guideline on Similar Biological Medicinal Products*. London, UK: European Medicines Agency; 2014.
3. FDA. *Guidance for Industry—Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*. Silver Spring, Maryland: Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), the United States Food and Drug Administration (FDA); 2015.
4. FDA. *Guidance for Industry - Statistical Approaches to Evaluate Analytical Similarity*. Silver Spring, Maryland: The United States Food and Drug Administration; 2017.
5. FDA. *Guidance for Industry - Considerations in Demonstrating Interchangeability with a Reference Product*. Silver Spring, Maryland: The United States Food and Drug Administration; 2019.
6. Jung EH, Sarpatwari A, Kesselheim AS, Sinha MS. FDA and EMA biosimilar approvals. *J Gen Intern Med*. 2020;35(6):1908–1910. doi:10.1007/s11606-019-05408-6
7. Tu CL, Wang YL, Hu TM, Hsu LF. Analysis of pharmacokinetic and pharmacodynamic parameters in EU-versus US-licensed reference biological products: are in vivo bridging studies justified for biosimilar development? *BioDrugs*. 2019;33(4):437–446. doi:10.1007/s40259-019-00357-2
8. Webster CJ, Woollett GR. A 'global reference'comparator for biosimilar development. *BioDrugs*. 2017;31(4):279–286. doi:10.1007/s40259-017-0227-4
9. Lim S. Overview of the regulatory framework and FDA's guidance for the development and approval of biosimilar and interchangeable products in the US. In: Slides for the July 13, 2017 Meeting of the Oncologic Drugs Advisory Committee (ODAC); Silver Spring, MD. Available from: https://www.fda.gov/drugs/biosimilars/fda-webinar-overview-regulatory-framework-and-fdas-guidance-development-and-approval-biosimilar-and. Accessed December 27, 2022.
10. Zheng J, Yin D, Yuan M, Chow SC. Simultaneous confidence interval methods for analytical similarity assessment. *J Biopharm Stat*. 2019;29(5):920–940. doi:10.1080/10543406.2019.1657142
11. Zheng C, Wang J, Zhao L. Testing bioequivalence for multiple formulations with power and sample size calculations. *Pharm Stat*. 2012;11(4):334–341. doi:10.1002/pst.1522
12. Chow SC. *Biosimilars: Design and Analysis of Follow-on Biologics*. New York: CRC Press; 2013.
13. Cho SH, Han S, Ghim JL, et al. A randomized, double-blind trial comparing the pharmacokinetics of CT-P16, a candidate bevacizumab biosimilar, with its reference product in healthy adult males. *BioDrugs*. 2019;33(2):173–181. doi:10.1007/s40259-019-00340-x
14. Desai K, Misra P, Kher S, Shah N. Clinical confirmation to demonstrate similarity for a biosimilar pegfilgrastim: a 3-way randomized equivalence study for a proposed biosimilar pegfilgrastim versus US-licensed and EU-approved reference products in breast cancer patients receiving myelosuppressive chemotherapy. *Exp Hematol Oncol*. 2018;7(1):1–11.