Sankhyā B : The Indian Journal of Statistics
https://doi.org/10.1007/s13571-024-00334-y

Improved Test Procedure and Sample Size Calculation for Assessing Similarity in Two-Group Comparative Studies with Heterogeneous Variances

Show-Li Jan
Chung Yuan Christian University, Taoyuan, Taiwan
Gwowen Shieh
National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Abstract

The two one-sided tests (TOST) method for mean equivalence or average equivalence has been extended to assessing similarity or switchability for individual equivalence in clinical trials. Tolerance interval procedures are available to establish similarity with respect to the proportion of the response differences covered by a prespecified threshold range. However, the extended TOST procedures based on tolerance intervals are potentially susceptible to the control of Type I errors. This article aims to present an exact approach with the specified Type I error probability for appraising similarity between two treatments in comparative studies with heterogeneous variances. Analytic examination and numerical comparison are conducted to clarify the utility of the suggested similarity test and the drawback of the current TOST procedures. To enhance the usefulness of the described exact method, the related power and sample size issues are also considered. Computer algorithms are provided to implement the proposed test procedure, power calculation, and sample size determination in similarity studies.

AMS (2000) subject classification. C12; C18; I10.
Keywords and phrases. Equivalence trials, method comparison, percentile, similarity test, tolerance interval.

1 Introduction

The conventional tests of significance focus primarily on the detection of difference between treatment effects. Alternatively, equivalence procedures provide a better approach to demonstrating agreement or compatibility for method comparisons in biological and medical sciences. The two one-sided tests (TOST) procedure of Schuirmann (1981) and Westlake (1981) is the most common method for claiming mean equivalence or average equivalence between two treatment groups. Despite the approximate nature, the TOST

Published online: 08 May 2024

procedure still maintains adequate control of Type I errors. The detection of mean equivalence by TOST is technically identical to assess whether the ordinary $100(1-2\alpha)\%$ two-sided confidence interval of mean difference lie within the designated equivalence bounds. Accordingly, the conceptual simplicity and computational ease facilitate general acceptance in practical equivalence problems. The technical discussion and fundamental review of different types of equivalence tests can be found in Berger and Hsu (1996) and Meyners (2012), respectively. Moreover, the concepts and techniques for the design and analysis of equivalence studies are described in Chow and Liu (2008), Hauschke et al. (2007), and Wellek (2010).

Various statistical principles and tools for measuring agreement were addressed in Barnhart et al. (2007), Choudhary and Nagaraja (2004, 2017), and Lin et al. (2012). Particularly, the TOST principle has been extended to evaluate similarity or switchability for individual equivalence in terms of the desired proportion of the measurement differences between two subjects. The basic concept and rationale of individual equivalence are given in Anderson (1993), Anderson and Hauck (1990), Hauck and Anderson (1992), Schall and Luus (1993), and Sheiner (1992). An important application of the similarity tests is to judge the individual bioequivalence between the test and reference formulations of a drug. Note that the one-sided confidence intervals of normal percentiles have a close link to the one-sided tolerance bounds of a normal distribution. Accordingly, tolerance interval technique is frequently used to evaluate the percentiles of measurement difference in similarity studies. General discussions of tolerance interval estimation are available in Krishnamoorthy and Mathew (2009) and Meeker et al. (2017). Consequently, the rejection regions of extended TOST tests are constructed with the tolerance limits for the designated proportions of normal distributions. The similarity problem is further complicated by the potentially unequal variances of the two treatment groups. Similar to the renowned Behrens-Fisher problem, approximate degrees-of-freedom techniques are often described to circumvent the inference issues under variance heterogeneity. Related heterogeneous TOST for mean equivalence are presented in Dannenberg et al. (1994), Dette and Munk (1997), Jan and Shieh (2017).

Several TOST procedures based on tolerance intervals were described in Chen and Hsiao (2020), and Dong et al. (2014). These TOST procedures declare similarity when the confidence limits for the percentiles of response differences are contained in the specified thresholds. Despite the obtained critical regions have a connection to the tolerance intervals, the TOST procedures are not constructed with respect to the principals of hypothesis testing.

Notably, Berger and Hsu (1996) noted that the incorrect association between the size-α tests with the $100(1-2\alpha)\%$ confidence sets is rather confusing and should be deemphasized. The overall message is that statistically sound techniques should be adopted to derive a test with the specified Type I error rate. Moreover, Berger and Hsu (1996) cautioned that there is no general guarantee that a TOST procedure in terms of a $100(1-2\alpha)\%$ confidence set will result in a size-α test. Related problems were also demonstrated in Shieh (2020, 2022) for evaluating agreement between two methods of quantitative measurements. The prescribed TOST methods for assessing similarity are intrinsically vulnerable to the ultimate problem of Type I error control. It is sensible to consider a proper test with the desired Type I error rate.

This article aims to describe an improved approach to establishing similarity between two treatments in comparative studies. The critical values are computed to meet the specified Type I error rate under the boundary parameter configurations of the null hypothesis. The proposed procedure declares similarity when the critical interval for the central proportion of measurement differences is within the designated threshold bounds. To explicate the relative behavior in Type I error control, simulation studies were conducted to examine and compare the rejection rates of the proposed approach and the TOST procedures. Moreover, power and sample size calculations of the suggested test are also described and evaluated. A real biosimilarity example of biological and reference products is presented to demonstrate the proposed techniques and computer algorithms for critical value, statistical power, and sample size calculations. The developed software programs are available as supplemental material.

2 The Proposed Similarity Test

Consider independent random samples from two normal populations:

$$X_{ij} \sim N(\mu_i, \sigma_i^2), \tag{1}$$

where μ_i and σ_i^2 are unknown parameters, $j = 1...N_i$, and $i = 1$ and 2. To establish the similarity between two treatment groups, the central portion of the difference between the individual measurements $X_{1j} - X_{2j'}$ needs to lie within a reasonable range around zero. The $100 \cdot p$th percentile of the distribution $N(\mu_D, \sigma_D^2)$ of $X_{1j} - X_{2j'}$ is denoted by

$$\theta_p = \mu_D + z_p \sigma_D, \tag{2}$$

where $\mu_D = \mu_1 - \mu_2$, $\sigma_D^2 = \sigma_1^2 + \sigma_2^2$, z_p is the the $100 \cdot p$th percentile of the standard normal distribution $N(0,1)$, and $0 < p < 1$. The null and alternative hypotheses of the similarity test are expressed as

$$H_0 : \theta_{1-p} \leq \Delta_L \text{ or } \Delta_U \leq \theta_p \text{ versus } H_1 : \Delta_L < \theta_{1-p} \text{ and } \theta_p < \Delta_U, \tag{3}$$

where $p > 0.5$ and the two designated constants Δ_L and Δ_U represent the lower and upper thresholds of the percentile range for declaring similarity between two groups. The alternative hypothesis indicates that there is at least $p^* = 2p - 1$ central proportion of the distribution $N(\mu_D, \sigma_D^2)$ in the range (Δ_L, Δ_U). Because of $\Delta_L < \theta_{1-p}$ and $\theta_p < \Delta_U$, the coverage probability $\Phi\{(\Delta_U - \mu_D)/\sigma_D\} - \Phi\{(\Delta_L - \mu_D)/\sigma_D\} > p^*$ where Φ is the cumulative density function of the standard normal distribution.

Within the framework of the Behren-Fisher problem, the approximate degrees-of-freedom procedure of Welch (1938) is commonly recommended as an alternative to the usual t test for mean comparison. The well-established Welch t statistic is of the form

$$T = \frac{D}{S_{DN}}, \tag{4}$$

where $D = \bar{X}_1 - \bar{X}_2$, $\bar{X}_1 = \sum_{j=1}^{N_1} X_{1j}/N_1$, $\bar{X}_2 = \sum_{j=1}^{N_2} X_{2j}/N_2$, $S_{DN}^2 = S_1^2/N_1 + S_2^2/N_2$, $S_1^2 = \sum_{j=1}^{N_1}(X_{1j} - \bar{X}_1)^2/\kappa_1$, $S_2^2 = \sum_{j=1}^{N_2}(X_{2j} - \bar{X}_2)^2/\kappa_2$, $\kappa_1 = N_1 - 1$, $\kappa_2 = N_2 - 1$. With the same theoretical arguments and analytic derivations in Welch (1938), it can be shown that the statistic T has the general approximate distribution

$$T \dot{\sim} t(\nu, \lambda), \tag{5}$$

where $t(\nu, \lambda)$ is a noncentral t distribution with degrees of freedom ν, and noncentrality parameter $\lambda = \mu_D/\sigma_{DN}$, $\sigma_{DN}^2 = \sigma_1^2/N_1 + \sigma_2^2/N_2$, and

$$\nu = \frac{(\sigma_1^2/N_1 + \sigma_2^2/N_2)^2}{(\sigma_1^2/N_1)^2/(N_1 - 1) + (\sigma_2^2/N_2)^2/(N_2 - 1)}.$$

In view of the desirable properties and practical applications of the T statistic, an extended Welch procedure is proposed here for similarity assessment.

For the equidistant range of (θ_{1-p}, θ_p) around the mean difference μ_D, the suggested exact rejection region for declaring similarity is of the form

$$\text{EXAT} = \{\Delta_L < \hat{\theta}_{EL} \text{ and } \hat{\theta}_{EU} < \Delta_U\}, \tag{6}$$

where $\hat{\theta}_{EL} = D - \tau_E S_{DN}$, $\hat{\theta}_{EU} = D + \tau_E S_{DN}$, and the quantity τ_E is designated to control the Type I error rate so that $\sup_{H_0} P\{\Delta_L \leq \hat{\theta}_{EL}$ and $\hat{\theta}_{EU} \leq \Delta_U\} = \alpha$. It is important to note that the supremum $\sup_{H_0} P\{\Delta_L \leq \hat{\theta}_{EL}$ and $\hat{\theta}_{EU} \leq \Delta_U\}$ is attained when the two percentiles coincide the boundary values $\{\theta_{1-p}, \theta_p\} = \{\Delta_L, \Delta_U\}$ or alternatively, $\mu_D = (\Delta_U + \Delta_L)/2$ and $\sigma_D = (\Delta_U - \Delta_L)/(2z_p)$. Thus, the actual magnitude of τ_E is determined by

$$\sup_{\Theta} P\{\theta_{1-p} \leq \hat{\theta}_{EL} \text{ and } \hat{\theta}_{EU} \leq \theta_p\} = \alpha \tag{7}$$

with respect to the boundary set of null parameters $\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) | \mu_D = (\Delta_U + \Delta_L)/2$ and $\sigma_D = (\Delta_U - \Delta_L)/(2z_p)\}$.

With the model assumption given in Eq. (1), it is evident that $Z = (D - \mu_D)/\sigma_{DN} \sim N(0,1)$ and $K = \kappa_1 S_1^2/\sigma_1^2 + \kappa_2 S_2^2/\sigma_2^2 \sim \chi^2(\kappa)$ where $\kappa = N_1 + N_2 - 2$ and $B = \{\kappa_1 S_1^2/\sigma_1^2\}/K \sim Beta\{\kappa_1/2, \kappa_2/2\}$. The random variables Z, K, and B are mutually independent. Moreover, the sample variance S_{DN}^2 of the sample mean difference D can be written as $S_{DN}^2 = K \cdot G$ where $G = (\sigma_1^2/N_1)(B/\kappa_1) + (\sigma_2^2/N_2)\{(1 - B)/\kappa_2\}$. Note that G is a function of the random variable B. The probability evaluation in Eq. (7) can be rewritten as

$$\sup_{\Theta} E_B E_K[P(-Z_0 < Z < Z_0)] = \sup_{\Theta} E_B E_K[2\Phi(Z_0) - 1] = \alpha, \tag{8}$$

where $Z_0 = \{z_p \sigma_D - \tau_E (K_0 G)^{1/2}\}/\sigma_{DN}$, $K_0 = \min\{K, (z_p^2 \sigma_D^2)/(\tau_E^2 G)\}$, and the expectations E_B and E_K are taken with respect to the distributions of B and K, respectively. It is vital to emphasize that the mean difference μ_D is irrelevant to the Type I error calculation and the quantity Z_0 can be simplified as a function of the variance ratio $\omega = \sigma_1^2/\sigma_2^2$. With the given values of variances (σ_1^2, σ_2^2) and other configurations, the particular critical value $\tau_E(\sigma_1^2, \sigma_2^2)$ that meets the equality $E_B E_K[2\Phi(Z_0) - 1] = \alpha$ can be determined with an iterative algorithm.

Note that the critical value $\tau_E(\sigma_1^2, \sigma_2^2)$ to attain the equality $E_B E_K[2\Phi(Z_0) - 1] = \alpha$ varies with the specified variance components (σ_1^2, σ_2^2) in which the sum of the two variance components $\sigma_D^2 = (\Delta_U - \Delta_L)^2/(4z_p^2)$.

Thus, the optimal critical value τ_E is the maximum of all critical values $\tau_E(\sigma_1^2, \sigma_2^2)$ correspond to the set of variance combinations $\{(\sigma_1^2, \sigma_2^2)|\sigma_D^2 = (\Delta_U - \Delta_L)^2/(4z_p^2)\}$. Due to the complexity nature, it requires a searching process to find the right solution. Detailed numerical investigations showed that the resulting values $\tau_E(\sigma_1^2, \sigma_2^2)$ has a U-shape form for σ_1^2 in $[0, (\Delta_U - \Delta_L)^2/(4z_p^2)]$. Therefore, the optimal critical value is the maximum of the two extremes as $\tau_E = \max\{\tau_E(0, \sigma_D^2), \tau_E(\sigma_D^2, 0)\}$. It is constructive to note when σ_D^2 is fixed that the variance $\sigma_{DN}^2 = \sigma_1^2/N_1 + \sigma_2^2/N_2 = \sigma_D^2/N_1$ if $N_1 = N_2$. Also, σ_{DN}^2 has a minimum $\min(\sigma_{DN}^2) = \sigma_D^2/N_1$ for $(\sigma_1^2, \sigma_2^2) = (\sigma_D^2, 0)$ if $N_1 > N_2$, and $\min(\sigma_{DN}^2) = \sigma_D^2/N_2$ for $(\sigma_1^2, \sigma_2^2) = (0, \sigma_D^2)$ if $N_1 < N_2$.

Following the prescribed results, the suggested agreement test rejects the null hypothesis if

$$\tau_E < T_L \text{ and } T_U < -\tau_E, \tag{9}$$

where $T_L = (D - \Delta_L)/S_{DN}$ and $T_U = (D - \Delta_U)/S_{DN}$. Under the alternative hypothesis, it can be shown that the power function of the suggested similarity test is of the form

$$\Psi_E = P\{\Delta_L < \hat{\theta}_{EL} \text{ and } \hat{\theta}_{EU} < \Delta_U\} = E_B E_K[\Phi(Z_U) - \Phi(Z_L)], \tag{10}$$

where $Z_L = \{\Delta_L - \mu_D + \tau_E(K_E G)^{1/2}/\sigma_{DN}$, $Z_U = \{\Delta_U - \mu_D - \tau_E(K_E G)^{1/2}/\sigma_{DN}$, and $K_E = \min\{K, (\Delta_U - \Delta_L)^2/(4\tau_E^2 G)\}$. The power function Ψ_E can be utilized to compute the minimal sample sizes for achieving the nominal power under the designated model configurations in planning research studies. The computations of the critical value and statistical power of the described extended Welch procedure can be readily conducted with the beta, chi-square, and normal probability functions in common statistical packages as shown in the supplementary materials.

3 TOST Procedures

The TOST procedure of Schuirmann (1981) and Westlake (1981) is widely used for evaluating mean equivalence or average equivalence between two treatment groups. To demonstrate comparability between two treatment means for the TOST test, it is statistically identical to examine whether the ordinary 100(1 - 2α)% equal-tailed confidence interval of mean difference is entirely within the equivalence bounds for declaring equivalence. The same notion was extended to individual equivalence assessment for interchangeability or biosimilarity in Chen and Hsiao (2020), and Dong et al. (2014).

3.1 The Dong, Tsong and Shen Procedure With the approximate degrees-of-freedom of the Welch (1938) statistic, it can be shown that the approximate lower confidence limit of a $100(1-2\alpha)\%$ equal-tailed confidence interval of θ_{1-p} is

$$\hat{\theta}_{WL} = D - \tau_W S_{DN}, \tag{11}$$

where $\tau_W = t_{1-\alpha}(\hat{\nu}, z_p H)$ is the $100(1-\alpha)\%$th percentile of a noncentral t distribution $t(\hat{\nu}, z_p H)$ with degrees of freedom $\hat{\nu}$ and noncentrality parameter $z_p H$ with $H^2 = S_D^2/S_{DN}^2$, $S_D^2 = S_1^2 + S_2^2$, and

$$\hat{\nu} = \frac{(S_1^2/N_1 + S_2^2/N_2)^2}{(S_1^2/N_1)^2/(N_1-1) + (S_2^2/N_2)^2/(N_2-1)}.$$

Also, the upper confidence limit of a $100(1-2\alpha)\%$ equal-tailed confidence interval of θ_p can be approximated by

$$\hat{\theta}_{WU} = D + \tau_W S_{DN}. \tag{12}$$

The resulting rejection region of the Welch-type TOST procedure in Dong et al. (2014, Section 2.2) is

$$\text{TOSTW} = \{\Delta_L < \hat{\theta}_{WL} \text{ and } \hat{\theta}_{WU} < \Delta_U\}. \tag{13}$$

Note that the formulations in Section 2.2 of Dong et al. (2014) have different notation and the critical value τ_W was denoted by $t_{1-\alpha}(\nu, z_p\eta)$ where $\eta^2 = \sigma_D^2/\sigma_{DN}^2$. It is a common practice to apply the substitution of (S_D^2, S_{DN}^2) for $(\sigma_D^2, \sigma_{DN}^2)$ in ν and η for data analysis. Evidently, the underlying properties of the approximation are somehow affected by the direct replacement.

3.2 The Chen and Hsiao Procedure To construct approximate one-sided tolerance limits for the difference of two independent normal variables, Hall (1984) suggested that the lower confidence limit of a $100(1-2\alpha)\%$ equal-tailed confidence interval of θ_{1-p} is

$$\hat{\theta}_{H1L} = D - \tau_{H1} S_D, \tag{14}$$

where $\tau_{H1} = t_{1-\alpha}(\hat{\nu}_1, z_p H_1)/H_1$, $\hat{\nu}_1 = (A_{11}^2/\kappa_1 + A_{12}^2/\kappa_2)^{-1}$, $A_{11} = a_1 S_1^2/(a_1 S_1^2 + S_2^2)$, $A_{12} = S_2^2/(a_1 S_1^2 + S_2^2)$, $H_1^2 = (a_1 S_1^2 + S_2^2)/(a_1 S_1^2/N_1 + S_2^2/N_2)$, and $a_1 = (N_2 - 3)/(N_2 - 1)$. On the other hand, the upper confidence limit of a $100(1-2\alpha)\%$ equal-tailed confidence interval of θ_p can be expressed as

$$\hat{\theta}_{H1U} = D + \tau_{H1} S_D. \tag{15}$$

Due to the undesirable features in the tolerance interval estimation of Hall (1984), Guo and Krishnamoorthy (2004) suggested that better estimates of the prescribed lower confidence limit of θ_{1-p} and upper confidence limit of θ_p can be obtained by

$$\hat{\theta}_{HL} = D - \tau_H S_D \text{ and } \hat{\theta}_{HU} = D + \tau_H S_D, \tag{16}$$

respectively, where $\tau_H = \max(\tau_{H1}, \tau_{H2})$, $\tau_{H2} = t_{1-\alpha}(\hat{\nu}_2, z_p H_2)/H_2$, $\hat{\nu}_2 = (A_{21}^2/\kappa_1 + A_{22}^2/\kappa_2)^{-1}$, $A_{21} = S_1^2/(S_1^2 + a_2 S_2^2)$, $A_{22} = a_2 S_2^2/(S_1^2 + a_2 S_2^2)$, $H_2^2 = (S_1^2 + a_2 S_2^2)/(S_1^2/N_1 + a_2 S_2^2/N_2)$, and $a_2 = (N_1 - 3)/(N_1 - 1)$. Note that the two percentiles τ_{H1} and τ_{H2} have parallel forms and are functions of a_1 and a_2, respectively. The consideration of the two quantities a_1 and a_2 for H_1 and H_2 is due to the fact that they yield unbiased estimation of the two different variance ratios $E[a_1 S_1^2/S_2^2] = \sigma_1^2/\sigma_2^2$ and $E[a_2 S_2^2/S_1^2] = \sigma_2^2/\sigma_1^2$. As a direct extension of Guo and Krishnamoorthy (2004), Chen and Hsiao (2020) considered the adapted rejection region for assessing similarity

$$\text{TOSTH} = \{\Delta_L < \hat{\theta}_{HL} \text{ and } \hat{\theta}_{HU} < \Delta_U\}. \tag{17}$$

3.3 Critical Values Note that the critical values τ_W and τ_H of the two TOST procedures are functions of the sample variances (S_1^2, S_2^2). Thus, the actual values of τ_W and τ_H presumably differ from sample to sample. In contrast, the critical value τ_E of the suggested approach is completely determined by the designated bounds (Δ_L, Δ_U) and does not depend on observed measurements. It should be emphasized that the estimated bounds $(\hat{\theta}_{EL}, \hat{\theta}_{EU})$, $(\hat{\theta}_{WL}, \hat{\theta}_{WU})$ and $(\hat{\theta}_{HL}, \hat{\theta}_{HU})$ for the prescribed rejection regions are all equidistant around the sample mean difference. The null hypothesis is rejected if such an interval is contained within the designated bounds (Δ_L, Δ_U). Accordingly, a narrower interval is more likely to reject the null hypothesis and to claim similarity between the two treatments.

When $p^* = 0.80$, $p = 0.90$, $\Delta_L = z_{0.10} = -1.2816$, $\Delta_U = z_{0.90} = 1.2816$, $N_1 = 10$, $N_2 = 20$, and $\alpha = 0.05$, it can be shown that the critical value is $\tau_E = 7.0605$ when the boundary parameter settings are $\mu_D = 0$ and $(\sigma_1^2, \sigma_2^2) = (0, 1)$. With the population variances $(\sigma_1^2, \sigma_2^2) = (0, 1)$, the sample variances may be $(S_1^2, S_2^2) = (0.0001, 0.9999)$ and then, the critical values of the three TOST procedures are $\tau_W = 8.6124$ and $\tau_H = 1.9260$. In this case of $S_D^2 = 1$ and $S_{DN}^2 = 0.0500$, the half-width of the intervals $(\hat{\theta}_{EL}, \hat{\theta}_{EU})$, $(\hat{\theta}_{WL}, \hat{\theta}_{WU})$ and $(\hat{\theta}_{HL}, \hat{\theta}_{HU})$ can be computed as $\tau_E S_{DN} = 1.5789$, $\tau_W S_{DN} = 1.9259$, and $\tau_H S_D = 1.9260$, respectively. Alternatively, when $(S_1^2, S_2^2) = (0.0020, 2.0000)$ or $S_D^2 = 2.0020$ and $S_{DN}^2 = 0.1002$, the critical values

change into $\tau_W = 8.6047$ and $\tau_H = 1.9256$. The corresponding half-widths for the intervals $(\hat{\theta}_{WL}, \hat{\theta}_{WU})$ and $(\hat{\theta}_{HL}, \hat{\theta}_{HU})$ are $\tau_W S_{DN} = 2.7238$ and $\tau_H S_D = 2.7246$, respectively. Also, the critical value remains as $\tau_E = 7.0605$ and the half-width of the critical interval $(\hat{\theta}_{EL}, \hat{\theta}_{EU})$ of the exact method is $\tau_E S_{DN} = 2.2350$.

These numerical results suggest that the paired bounds $(\hat{\theta}_{WL}, \hat{\theta}_{WU})$ and $(\hat{\theta}_{HL}, \hat{\theta}_{HU})$ of the TOST procedures are nearly identical, whereas the bounds $(\hat{\theta}_{EL}, \hat{\theta}_{EU})$ of the proposed approach apparently have a smaller interval. Because the exact method is constructed to have the desired control of Type I error rate, the comparisons reveal the TOST methods may reject the null hypothesis less often than the nominal level and tend to be conservative test procedures. The performance of these similarity tests will be further examined in the subsequent numerical investigations.

4 Type I Errors

Numerical results were presented in Chen and Hsiao (2020), and Dong et al. (2014) to justify the TOST procedures for assessing similarity. However, the Type I error appraisals of the TOST methods were not examined with respect to the supremum of the boundary set of null parameters Θ defined in Eq. (7). A proper and thorough evaluation is required to demonstrate the underlying behavior of the similarity tests. Accordingly, simulation study was conducted to inspect their Type I error performance under a variety of model configurations.

To elucidate the potential discrepancy between the suggested approach and the TOST procedures, the numerical investigations cover the central proportion $p^* = 0.80$, 0.90, and 0.95. For ease of illustration, the mean and variance of the null distribution $N(\mu_{D0}, \sigma^2_{D0})$ for the measurement difference $X_{1j} - X_{2j'}$ is chosen as $\mu_{D0} = 0$ and $\sigma^2_{D0} = 1$. Accordingly, the designated thresholds (Δ_L, Δ_U) are determined by $\Delta_L = \mu_{D0} - z_p\sigma_{D0}$ and $\Delta_U = \mu_{D0} + z_p\sigma_{D0}$. The resulting similarity bounds are (Δ_L, Δ_U) = (-1.2816, 1.2816), (-1.6449, 1.6449), and (-1.9600, 1.9600) for $p = 0.90$, 0.95, and 0.975, respectively. Four sets of sample sizes are considered: (N_1, N_2) = (10, 20), (20, 50), (50, 100), and (100, 200). Through the empirical examination, the significance level is fixed as $\alpha = 0.05$.

The simulated Type I error rates of the agreement tests were computed via Monte Carlo simulation of 10,000 independent data sets. Under the model settings, the optimal critical value τ_E of the proposed procedure is attained when $(\sigma^2_1, \sigma^2_2) = (0, 1)$. To avoid computational ambiguity and maintain theoretical implication, the variance components are slightly modified as (σ^2_1, σ^2_2)

= (0.0001, 0.9999). For the four test procedures, the simulated Type I error rates were the proportion of the 10,000 replicates whose critical intervals $(\hat{\theta}_{EL}, \hat{\theta}_{EU})$, $(\hat{\theta}_{WL}, \hat{\theta}_{WU})$ and $(\hat{\theta}_{HL}, \hat{\theta}_{HU})$ were within the range of (Δ_L, Δ_U). Accordingly, the simulated Type I error probabilities under the four different sample sizes are summarized in Tables 1, 2, and 3 for the three central portions $p^* = 0.80$, 0.90, and 0.95, respectively. The adequacy of the contending procedures is determined by the difference between the simulated outcomes and the nominal level 0.05 as shown in the tables.

The results in Tables 1-3 showed that the simulated Type I error rates of the suggested approach are rather close to the nominal value 0.05. Evidently, the proposed similarity test has excellent control of Type I errors for the model configurations considered here. In contrast, the performance of the TOST procedures is highly disturbing. Due to the supremum consideration, the simulated type I error probabilities of the TOST methods of TOSTW and TOSTH have the identical values. Specifically, the resulting simulated Type I error rates are within the ranges of [0.0024, 0.0054], [0.0063, 0.0105], and [0.0096, 0.0122] in Tables 1-3, respectively. The discrepancy between the simulated alpha and nominal alpha only marginally decreased with larger sample sizes. The small rejection rates suggest that the TOST procedures are overly conservative. Unlike the demonstration and evaluation presented in Chen and Hsiao (2020), and Dong et al. (2014), these findings reveal that the TOST procedures do not have adequate Type I error control and cannot be recommended for similarity assessments.

5 Power and Sample Size Calculations

Power and sample size calculations are crucial elements in planning research designs. The related power and sample size problems for equivalence and agreement tests were addressed in Shieh (2016, 2020, 2022). It is of practical concern to explicate the power and sample issues of the current similarity test under variance heterogeneity. Accordingly, simulation studies were conducted to demonstrate the accuracy of derived power function and the usefulness of accompanying computer algorithm in sample size determinations. Sample size planning requires detailed specifications of Type I error rate α, nominal power $1-\beta$, equivalence bounds (Δ_L, Δ_U), null central portion p^*, and the alternative settings include the mean values (μ_1, μ_2), error variances (σ_1^2, σ_2^2), and sample size allocation ratio $r = N_2/N_1$. Note that the designated parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ are chosen such that $\Delta_L < \theta_{1-p}$ and $\theta_p < \Delta_U$ under the alternative distribution $N(\mu_D, \sigma_D^2)$.

Table 1: The simulated Type I error rates of similarity tests for central proportion $p^* = 0.80$, equivalence bounds $(\Delta_L, \Delta_U) = (-1.2816, 1.2816)$, and the significance level $\alpha = 0.05$

	Sample sizes (N_1, N_2)							
	(10, 20)		(20, 50)		(50, 100)		(100, 200)	
	Simulated difference*		Simulated difference		Simulated difference		Simulated difference	
Test procedure	α		α		α		α	
The proposed approach	0.0476	−0.0024	0.0507	0.0007	0.0504	0.0004	0.0487	−0.0013
Dong et al. (2014)	0.0054	−0.0446	0.0035	−0.0465	0.0032	−0.0468	0.0024	−0.0476
Chen and Hsiao (2020)	0.0054	−0.0446	0.0035	−0.0465	0.0032	−0.0468	0.0024	−0.0476

Note: $\Delta_L = \mu_D - z_p\sigma_D$ and $\Delta_U = \mu_D + z_p\sigma_D$ where $\mu_D = 0$, $\sigma_D^2 = 1$, $p = 0.90$, $z_p = 1.2816$

*Difference refers to the difference between the simulated α and the nominal level 0.05

Table 2: The simulated Type I error rates of similarity tests for central proportion $p^* = 0.90$, equivalence bounds $(\Delta_L, \Delta_U) = (-1.6449, 1.6449)$, and the significance level $\alpha = 0.05$

	Sample sizes (N_1, N_2)							
	(10, 20)		(20, 50)		(50, 100)		(100, 200)	
	Simulated difference*		Simulated difference		Simulated difference		Simulated difference	
Test procedure	α		α		α		α	
The proposed approach	0.0486	−0.0014	0.0503	0.0003	0.0520	0.0020	0.0491	−0.0009
Dong et al. (2014)	0.0105	−0.0395	0.0075	−0.0425	0.0070	−0.0430	0.0063	−0.0437
Chen and Hsiao (2020)	0.0105	−0.0395	0.0075	−0.0425	0.0070	−0.0430	0.0063	−0.0437

Note: $\Delta_L = \mu_D - z_p\sigma_D$ and $\Delta_U = \mu_D + z_p\sigma_D$ where $\mu_D = 0$, $\sigma_D^2 = 1$, $p = 0.95$, $z_p = 1.6449$

*Difference refers to the difference between the simulated α and the nominal level 0.05

Table 3: The simulated Type I error rates of similarity tests for central proportion $p^* = 0.95$, equivalence bounds $(\Delta_L, \Delta_U) = (-1.9600, 1.9600)$, and the significance level $\alpha = 0.05$

	Sample sizes (N_1, N_2)							
	(10, 20)		(20, 50)		(50, 100)		(100, 200)	
	Simulated difference*		Simulated difference		Simulated difference		Simulated difference	
Test procedure	α		α		α		α	
The proposed approach	0.0505	0.0005	0.0502	0.0002	0.0499	−0.0001	0.0527	0.0027
Dong et al. (2014)	0.0122	−0.0378	0.0107	−0.0393	0.0096	−0.0404	0.0097	−0.0403
Chen and Hsiao (2020)	0.0122	−0.0378	0.0107	−0.0393	0.0096	−0.0404	0.0097	−0.0403

Note: $\Delta_L = \mu_D - z_p\sigma_D$ and $\Delta_U = \mu_D + z_p\sigma_D$ where $\mu_D = 0$, $\sigma_D^2 = 1$, $p = 0.975$, $z_p = 1.9600$

*Difference refers to the difference between the simulated α and the nominal level 0.05

In the following numerical investigations, two central portions are considered: $p^* = 0.90$ and 0.95 ($p = 0.95$ and 0.975). The corresponding threshold bounds are (Δ_L, Δ_U) = (-1.6449, 1.6449), and (-1.9600, 1.9600), respectively. For the alternative distribution, the treatment means are (μ_1, μ_2) = (0, 0), (0.05, 0), and (0.10, 0). Also, three pairs of error variances are evaluated: $(\sigma_1^2, \sigma_2^2) = \{(1/3)\sigma_D^2, (2/3)\sigma_D^2\}$ for $\sigma_D^2 = 0.6$, 0.7 and 0.8. With the selected configuration, the minimum total sample size $N_T = N_1 + N_2$ is computed for the balanced design $r = 1$ ($N_1 = N_2$), significance level $\alpha = 0.05$, and nominal power $1-\beta = 0.8$. The estimated sample sizes and estimated power levels are summarized in Table 4 for the combined 18 cases. It can be seen from the results in Table 4 that the total sample sizes cover a wide range of values. The smallest sample size is 96 under the settings of $p^* = 0.95$, $\mu_D = 0$, and $\sigma_D^2 = 0.6$. In contrast, the particular scenario of $p^* = 0.90$, $\mu_D = 0.10$, and $\sigma_D^2 = 0.8$ give the largest sample size 1838. Evidently, these vital configurations impose unique and distinct impact in power and sample size calculations.

Moreover, simulation study was conducted to justify the accuracy of the proposed power and sample size procedures. Under the prescribed model configurations, the simulated power of the proposed similarity test was computed via Monte Carlo simulation of 10,000 independent data sets. The simulated power and the difference between the simulated power and estimated power are also summarized in Table 4. The small differences show that the simulated power is almost identical to the estimated power. Thus, the suggested power and sample size algorithms are accurate for general use. However, the proposed techniques are not currently available in statistical packages. Computer algorithms are developed to facilitate the application of the recommended approaches for similarity studies. The achieved power levels and estimated sample sizes can provide useful guidance about the meaning and influence of the vital factors in the intended research.

6 An Application

To further exemplify the utility of the suggested techniques and accompanying programs, a biosimilarity example in Chen and Hsiao (2020) is presented and extended for the suggested assessments of similarity, power analysis, and sample size determination.

Chen and Hsiao (2020) discussed a problem of appraising the biosimilarity of the biological Epoetin Hospira with the reference product Epogen/Procrit as documented in FDA (2017). The primary endpoint being examined in Chen and Hsiao (2020) is the mean weekly dosage per kilogram of body

Table 4: Estimated sample size, estimated power, and simulated power of the proposed similarity test for balanced design $N_1 = N_2, \{\sigma_1^2, \sigma_2^2\} = \{(1/3)\sigma_D^2, (2/3)\sigma_D^2\}$, the nominal power 0.80, and the significance level $\alpha = 0.05$

Null proportion p^*	Equivalence bounds (Δ_L, Δ_U)	Mean μ_D	Variance σ_D^2	Total sample size N_T	Simulated power	Estimated power	Difference
0.90	(−1.6449, 1.6449)	0	0.6	98	0.8021	0.8011	0.0010
			0.7	202	0.8030	0.8023	0.0007
			0.8	518	0.7974	0.8004	−0.0030
		0.05	0.6	104	0.8014	0.8021	−0.0007
			0.7	226	0.8068	0.8018	0.0050
			0.8	694	0.7997	0.8002	−0.0005
		0.10	0.6	126	0.8054	0.8031	0.0023
			0.7	332	0.7998	0.8002	−0.0004
			0.8	1838	0.7998	0.8002	−0.0004
0.95	(−1.9600, 1.9600)	0	0.6	96	0.8092	0.8077	0.0015
			0.7	194	0.8019	0.8039	−0.0020
			0.8	492	0.8017	0.8004	0.0013
		0.05	0.6	100	0.8065	0.8057	0.0008
			0.7	210	0.8000	0.8004	−0.0004
			0.8	614	0.8029	0.8003	0.0026
		0.10	0.6	114	0.8049	0.8012	0.0037
			0.7	280	0.8039	0.8007	0.0032
			0.8	1236	0.8003	0.8004	−0.0001

weight during the last 4 weeks of the double-blind treatment period. The sample sizes, sample means, and sample variances of the groups of US Epogen/Procrit and Epoetin Hospira are (N_1, N_2) = (122, 124), $(\bar{X}_1, \bar{X}_2)$ = (81.9, 79.6), and (S_1^2, S_2^2) = (2329.8218, 2357.1904), respectively. With respect to the central portion $p^* = 0.90$, the reference bounds for the similarity study are chosen as (Δ_L, Δ_U) = (-157.29, 157.29). It can be shown that the mean difference $D = 2.3$, sample standard deviation $S_{DN} = 6.1730$, and critical value $\tau_E = 19.8063$ at the significance level α=0.05. The resulting critical region can readily be obtained as $(\hat{\theta}_{EL}, \hat{\theta}_{EU})$ = (-119.9654, 124.5654). It is clear that the interval falls within the designated thresholds (Δ_L, Δ_U). Thus, the null hypothesis $H_0 : \theta_{0.05} \leq$ -157.29 or 157.29 $\leq \theta_{0.95}$ is rejected and the result declares the biosimilarity properties between the two drug formulations of Epoetin Hospira and Epogen/Procrit.

Additional simulation study was conducted using the summary statistics of the prescribed biosimilarity study as the population means and variances where (μ_1, μ_2) = (81.9, 79.6) and (σ_1^2, σ_2^2) = (2329.8218, 2357.1904). With (N_1, N_2) = (122, 124), (Δ_L, Δ_U) = (-157.9, 157.29), $p^* = 0.90$, and $\alpha = 0.05$, the simulated Type I error rates of 10,000 replications are 0.0502, 0.0065 and 0.0065 for the EXAT, TOSTW and TOSTH procedures, respectively. These results are consistent with those reported in the prescribed simulation study of Type I errors under a wide variety of model configurations. For the purpose of conducting power calculation and sample size determination for similarity design, the minimum sample sizes to attain the nominal power 0.80 are $N_1(= N_2) = 11$, 29 and 123 for $p^* = 0.80, 0.90$ and 0.95 ($p = 0.90, 0.95$ and 0.975), respectively. The achieved power levels for the sample sizes 0.8345, 0.8075, and 0.8005 are marginally larger than the nominal level 0.80. Moreover, the corresponding sample sizes for the nominal power 0.90 are obtained as $N_1(= N_2) = 13, 37,$ and 163 with the estimated power levels 0.9053, 0.9026, and 0.9007 for $p^* = 0.80, 0.90$ and 0.95, respectively. The sharp increase of optimal sample size from $p^* = 0.90$ to 0.95 suggests that the common rule of thumb and simple linear interpolation are unlikely to account for such structural change and delicate balance in model configurations. A detailed and reliable procedure is essential to provide accurate power and sample size calculations. These exemplifying configurations are presented in the user specifications of the supplemental computer programs. Users can readily accommodate their own model specifications by specifying the chosen values in these statements.

7 Conclusions

This paper presents an improved test procedure for assessing similarity in two-group comparative studies. It is rigorously shown that the size of the suggested approach is exactly equal to the nominal Type I error probability. Alternatively, TOST extensions for establishing similarity have been considered in Dong et al. (2014), and Chen and Hsiao (2020), among others. Despite the existing technical arguments and empirical evidences for the contending TOST procedures, detailed numerical investigations reveal their underlying deficiency in Type I error control. Specifically, these TOST procedures based on tolerance intervals are excessively conservative. The results agree with the concern of Berger and Hsu (1996) that the practice of evaluating bioequivalence tests in terms of a $100(1-2\alpha)\%$ confidence intervals for average equivalence may not be sensible. Simulation studies were also conducted to justify the usefulness of the suggested power and sample size procedures for similarity analysis. Computer algorithms are presented to facilitate the implementation of the proposed similarity test, power calculation, and sample size determination.

Funding Open Access funding enabled and organized by National Yang Ming Chiao Tung University. This study was funded by Ministry of Science and Technology.

Data Availability The datasets generated during and/or analyzed during the current study are available in the cited references.

Declarations

Conflict of interests The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Anderson, S. (1993). Individual bioequivalence: A problem of switchability (with discussion). *Biopharmaceutical Reports*, 2(2), 1–11.

Anderson, S. & Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, 18, 259–273.

Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics*, 17(4), 529–569.

Berger, R. L. & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets (with discussion). *Statistical Science*, 11, 283–319.

Chen, C. & Hsiao, C. F. (2020). Use of tolerance intervals for assessing biosimilarity. *Statistics in Medicine*, 39(26), 3806–3822.

Choudhary, P. K. & Nagaraja, H. (2004). Measuring agreement in method comparison studies-a review. *Advances in ranking and selection, multiple comparisons, and reliability: Methodology and applications*, pp. 215–244.

Choudhary, P. K. & Nagaraja, H. N. (2017). *Measuring agreement: Models, methods, and applications*. John Wiley & Sons.

Chow, S.-C. & Liu, J. P. (2008). *Design and analysis of bioavailability and bioequivalence studies*. CRC press.

Dannenberg, O., Dette, H., & Munk, A. (1994). An extension of welch's approximate t-solution to comparative bioequivalence trials. *Biometrika*, 81(1), 91–101.

Dette, H. & Munk, A. (1997). Optimum allocation of treatments for Welch's test in equivalence assessment. *Biometrics*, 53, 1143–1150.

Dong, X., Tsong, Y., & Shen, M. (2014). Equivalence tests for interchangeability based on two one-sided probabilities. *Journal of Biopharmaceutical Statistics*, 24(6), 1332–1348.

FDA (2017). Advisory Committee Epoetin Hospira: A Proposed Biosimilar to Epogen/Procrit (Epoetin Alfa). *FDA Briefing Document: BLA 125545*.

Guo, H. & Krishnamoorthy, K. (2004). New approximate inferential methods for the reliability parameter in a stress–strength model: The normal case. *Communications in Statistics-Theory and Methods*, 33(7), 1715–1731.

Hall, I. J. (1984). Approximate one-sided tolerance limits for the difference or sum of two independent normal variates. *Journal of Quality Technology*, 16(1), 15–19.

Hauck, W. & Anderson, S. (1992). Types of bioequivalence and related statistical considerations. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, 30(5), 181–187.

Hauschke, D., Steinijans, V., & Pigeot, I. (2007). *Bioequivalence studies in drug development: Methods and applications*. John Wiley & Sons.

Jan, S. L. & Shieh, G. (2017). Optimal sample size determinations for the heteroscedastic two one-sided tests of mean equivalence: Design schemes and software implementations. *Journal of Educational and Behavioral Statistics*, 42(2), 145–165.

Krishnamoorthy, K. & Mathew, T. (2009). *Statistical tolerance regions: Theory, applications, and computation*. John Wiley & Sons.

Lin, L., Hedayat, A., & Wu, W. (2012). *Statistical tools for measuring agreement*. Springer.

Meeker, W. Q., Hahn, G. J., & Escobar, L. A. (2017). *Statistical intervals: A guide for practitioners and researchers*, volume 541. John Wiley & Sons.

Meyners, M. (2012). Equivalence tests–A review. *Food Quality and Preference*, 26(2), 231–245.

Schall, R. & Luus, H. G. (1993). On population and individual bioequivalence. *Statistics in medicine*, 12(12), 1109–1124.

Schuirmann, D. (1981). On hypothesis testing to determine if the mean of a normal-distribution is contained in a known interval. *Biometrics*, 37(3), 617–617.

Sheiner, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine*, 11(13), 1777–1788.

Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS ONE*, 11(9), e0162093.

Shieh, G. (2020). Assessing agreement between two methods of quantitative measurements: Exact test procedure and sample size calculation. *Statistics in Biopharmaceutical Research*, 12(3), 352–359.

Shieh, G. (2022). Assessing individual equivalence in parallel group and crossover designs: Exact test and sample size procedures. *PLoS ONE*, 17(5), e0269128.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4), 350–362.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Boca raton Florida: Chapman & Hal/CRC.

Westlake, W. J. (1981). Bioequivalence testing-a need to rethink. *Biometrics*, 37(3), 589–594.

Show-Li Jan
Department of Applied Mathematics, Chung Yuan Christian University, Taoyuan 32023, Taiwan
E-mail: sljan@cycu.edu.tw

Gwowen Shieh
Department of Management Science, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu 300093, Taiwan
E-mail: gwshieh@nycu.edu.tw

Paper received: 26 November 2023; accepted 22 April 2024.