RESEARCH ARTICLE

An evaluation roadmap for critical quality attributes from tier 1 in analytical similarity assessment

Kejian Wu[1,2☯], **Haitao Pan**[3☯], **Chen Li**[1], **Qingbo Zhao**[2], **Ling Wang**[1*], **Jielai Xia**[1*]

1 Department of Health Statistics, Fourth Military Medical University, Xi'an, Shannxi, China, **2** Department of Mathematics and Physics, Fourth Military Medical University, Xi'an, Shannxi, China, **3** Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America

☯ These authors contributed equally to this work.
* jielaixia@yahoo.com (JX); wl.medstat@gmail.com (LW)

OPEN ACCESS

Citation: Wu K, Pan H, Li C, Zhao Q, Wang L, Xia J (2018) An evaluation roadmap for critical quality attributes from tier 1 in analytical similarity assessment. PLoS ONE 13(12): e0208354. https://doi.org/10.1371/journal.pone.0208354

Editor: Yongshuai Jiang, Harbin Medical University, CHINA

Received: August 17, 2018

Accepted: November 15, 2018

Published: December 6, 2018

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This research was supported by the National major scientific and technological special project for significant new drugs development of China (2015ZX09501008-004) to JX, and the National Natural Science Foundation of China (81773553) to LW. Pan's research was supported by the American Lebanese and Syrian Associated Charities.

Abstract

Analytical similarity assessment of critical quality attributes (CQAs) serves as a foundation for the development of biosimilar products and facilitates an abbreviated subsequent clinical evaluation. In this study, we establish a statistical evaluation roadmap with statistical approaches for some selected CQAs from Tier 1, because they are most relevant to clinical outcomes and require the most rigorous statistical methods. In the roadmap, we incorporate 3 methods—ranking and tier assignment of quality attributes, the equivalence test, and the Mann–Whitney test for equivalence—that are important to determine analytical similarity between the reference and biosimilar products. For the equivalence test, we develop a power calculation formula based on the two one-sided tests procedure. Exact sample sizes can be numerically calculated. Then, we propose a flexible idea for selecting the number of reference lots (n_R) and the number of biosimilar lots (n_T) to adjust for serious unbalanced sample sizes. From results of extensive simulations under various parameter settings, we obtain a workable strategy to determine the optimum sample size combination (n_T, n_R) for the equivalence test of CQAs from Tier 1. R codes are provided to facilitate implementation of the roadmap and corresponding methods in practice.

Introduction

Biosimilars are biological products that are highly similar but not identical to their reference products, notwithstanding minor differences in clinically inactive components. Thus, biosimilars are close but not exact copies of biological products that are already on the market. With the expiration of patents on many innovative biological products, biosimilar products have received increasing attention from pharmaceutical companies such as Celltrion [1], Pfizer [2], and Sandoz [3] and from regulatory agencies such as the European Medicines Agency [4], United States Food and Drug Administration (FDA) [5, 6], World Health Organization [7], and China Food and Drug Administration [8]. Biosimilars can offer affordable treatment alternatives for diseases such as cancer and chronic inflammatory disorders.

Competing interests: The authors have declared that no competing interests exist.

It is important for biosimilar developers to understand how to demonstrate that the product is biosimilar to its reference product. FDA guidelines recommend a stepwise approach to generate data needed to demonstrate biosimilarity [5]. The stepwise approach is briefly summarized in the pyramid, as shown in Fig 1 proposed by Chow [9]. The stepwise approach starts with analytical studies of critical quality attributes (CQAs) that are relevant to clinical outcomes [10]. The shape of the pyramid signifies that fewer data are required in the clinical phase if adequate biosimilarity has been established in previous steps. For example, comprehensive analytical characterization was used to assess the analytical similarity between ABP 501 and 2 adalimumab products [11] and between ABP 215 and both United States–and European Union–sourced bevacizumab products [12].

Considering that there may be a large number of CQAs in practice, Chow [9] and Tsong et al. [10] proposed a statistical approach for demonstrating analytical similarity based on a tiered system that accounts for their criticality, for example, most (Tier 1), mild to moderate (Tier 2), and least (Tier 3) relevant to clinical outcomes. They also recommended the equivalence test of means for CQAs from Tier 1, the quality range approach for CQAs from Tier 2, and visual displays for CQAs from Tier 3. Since the most rigorous statistical method is required for CQAs from Tier 1, many statisticians have performed important pioneering

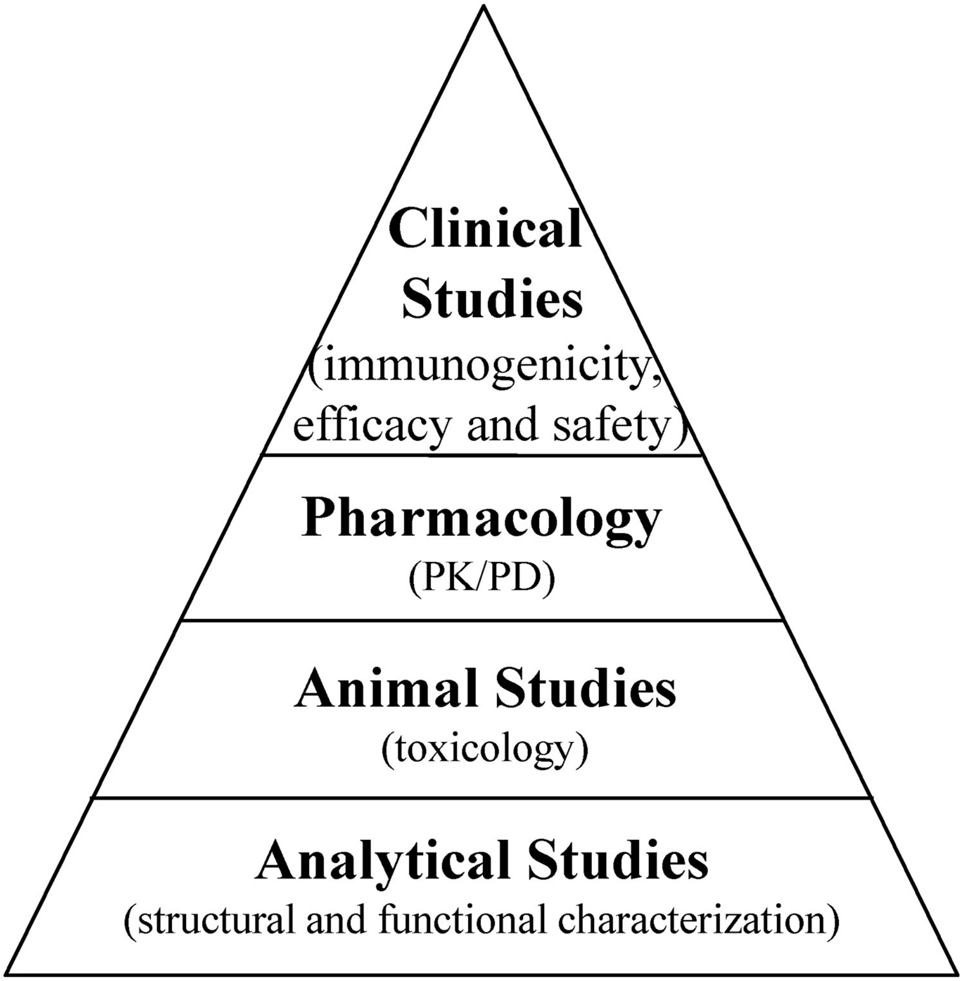
Clinical Studies
(immunogenicity, efficacy and safety)
Pharmacology
(PK/PD)
Animal Studies
(toxicology)
Analytical Studies
(structural and functional characterization)

Fig 1. Stepwise approach to assess biosimilarity. PK: pharmacokinetics; PD: pharmacodynamics.

https://doi.org/10.1371/journal.pone.0208354.g001 studies on CQAs from Tier 1. For example, Chow et al. discussed properties of the equivalence test [13], justification for margin [14], and sample size [15]. Tsong et al. provided details of the equivalence test [10]. Dong et al. proposed 2 sample size imbalance adjustment methods [16]. Other issues have been considered in Shen et al. [17], Burdick et al. [18], Dong et al. [19], Chen et al. [20], Liao et al. [21], and Wu et al. [22].

However, these studies have often focused on a particular statistical issue and have not developed a complete evaluation system for biosimilar developers, especially those conducting quality analytical tests. Therefore, in this study, we develop a statistical evaluation roadmap for some selected CQAs from Tier 1, focusing on both statistical methods and simplicity of implementation. The goal of our roadmap is to provide evaluation procedures to biosimilar developers in an accessible manner.

This paper is organized as follows. Section 2 introduces key factors in the evaluation roadmap: (i) the risk ranking and tier assignment of quality attributes (QAs), (ii) statistical considerations of equivalence test—power function and sample size required, and (iii) Mann–Whitney test for equivalence for seriously skewed analytical data. Section 3 presents a case study. Section 4 presents concluding remarks with discussions.

Methods

For analytical similarity assessment of a biosimilar, a comprehensive analytical characterization is performed to compare the proposed biosimilar and reference products. For physical/chemical characterization of products, we can obtain a large number of testing values of QAs by using state-of-the-art analytical methods. These QAs may include general properties, primary structure, higher-order structure, particles and aggregates, product-related substances and impurities, biological activity and forced thermal degradation, and so on. It is impractical to statistically compare all QAs to demonstrate biosimilarity. Thus, the identification of CQAs among QAs is an important first step in analytical similarity assessment, which is based on a thorough understanding of the potential for QAs to affect safety and efficacy. Thus, we first introduce a systematic scientific and risk-based approach to identify CQAs and assign their tiers. Second, we study statistical approaches for the equivalence test for some selected CQAs from Tier 1. Successful completion of these steps will ensure that there is sufficient evidence to demonstrate that a proposed biosimilar is highly similar to its reference product in analytical similarity assessment.

Ranking and tier assignment of quality attributes

To identify CQAs from a lot of QAs, we recommend the risk ranking and filtering approach developed by Roche/Genentech [23]. This approach focuses on drug safety and efficacy and incorporates 2 factors: impact and uncertainty of that impact. Impact is assigned on a 2- to 20-point scale that considers the known or potential effect of an attribute on 4 clinical performance categories: bioactivity, pharmacokinetics, immunogenicity, and safety. Uncertainty is based on the confidence that biosimilar developers have in the relevance of the information used in impact assessment. Uncertainty is assigned on a 1- to 7-point scale, with lower scores reflecting higher confidence. Then, the risk score of an attribute is generated by multiplying the 2 values of impact and uncertainty:

$$\text{Risk} = \text{Impact}(2 - 20) \times \text{Uncertainty}(1 - 7). \tag{1}$$

The highest risk score of the above 4 categories is used to categorize the QA as CQA or non-CQA. Then, 13 risk scores are selected as the cutoff. That is, attributes having risk scores greater than 13 in any single impact category are classified as CQAs. Alt et al. provide further details on the ranking and determination of CQAs and examples from monoclonal antibodies [23].

After many QAs are classified as CQAs, biosimilar developers need to determine the appropriate tier of CQAs. Tiers are assigned based on the risk score, and Tier 1 is reserved for the highest risk scores that have a direct impact on clinical outcomes. In addition to the highest risk scores, several other factors such as quantitative or qualitative data and the level of assays used for assessing attributes should also be considered [24]. Criticality and determination of tiering of CQAs are assessed mainly by biosimilar developers in the analytical characterization or biocharacterization team. In the following subsections, we propose statistical approaches for some selected CQAs from Tier 1 that are appropriate for the equivalence test.

Equivalence test for CQAs from Tier 1

We conduct the test for equivalence of means of selected CQAs from Tier 1 between the proposed biosimilar and reference products. Let T and R be the responses of a given CQA from Tier 1 for the biosimilar (or test) product and its reference product, respectively. Assuming that T and R follow a $N(\mu_T, \sigma_T^2)$ and $N(\mu_R, \sigma_R^2)$ distribution, where μ_T and μ_R are mean values, σ_T^2 and σ_R^2 are the variances, respectively. By using a parallel design, we test the following hypothesis:

$$H_0 : \mu_T - \mu_R \leq -\delta \text{ or } \mu_T - \mu_R \geq \delta \text{ vs. } H_a : -\delta < \mu_T - \mu_R < \delta, \tag{2}$$

where $\delta > 0$ is the equivalence margin. This type of test can be decomposed into Schuirmann's two one-sided tests [25], in which H_0 and H_a in (2) are tested separately by a one-sided test:

$$H_{01} : \mu_T - \mu_R \leq -\delta \text{ vs. } H_{a1} : \mu_T - \mu_R > -\delta \tag{3}$$

$$H_{02} : \mu_T - \mu_R \geq \delta \text{ vs. } H_{a2} : \mu_T - \mu_R < \delta. \tag{4}$$

We then reject H_{01} at the α level of significance in (3) if

$$T_L = \frac{(\bar{X}_T - \bar{X}_R) + \delta}{\sqrt{S_T^2/n_T + S_R^2/n_R}} > t_{\alpha,v} \tag{5}$$

and reject H_{02} in (4) if

$$T_U = \frac{(\bar{X}_T - \bar{X}_R) - \delta}{\sqrt{S_T^2/n_T + S_R^2/n_R}} < -t_{\alpha,v}, \tag{6}$$

where sample sizes n_T and n_R refer to the number of lots from the proposed biosimilar and the reference product required in the equivalence test, respectively. $\bar{X}_T$, $\bar{X}_R$ and S_T, S_R are the sample mean and standard deviation (SD) of the proposed biosimilar and the reference products, respectively. The symbol $t_{\alpha,v}$ is the α 100%th percentile of the t-distribution with the degrees of freedom approximated by Satterthwaite's approximation as $v = (S_T^2/n_T + S_R^2/n_R)^2/[S_T^4/[n_T^2(n_T - 1)] + S_R^4/[n_R^2(n_R - 1)]]$ [26].

The global null hypothesis H_0 in (2) is rejected with type I error α if both one-sided hypotheses (3) and (4) are rejected with type I error α. Thus, we conclude that there is sufficiently strong evidence to support statistical equivalence in means if both one-sided hypotheses H_{01} in (3) and H_{02} in (4) are rejected.

An alternative method to assess similarity between the 2 products is to use a two-sided confidence interval (CI) for $\mu_T - \mu_R$. We conclude that there is statistical equivalence in means if the 100(1 − 2α)% CI $\left(\bar{X}_T - \bar{X}_R - t_{\alpha,v} \cdot \sqrt{\frac{S_T^2}{n_T} + \frac{S_R^2}{n_R}}, \bar{X}_T - \bar{X}_R + t_{\alpha,v} \cdot \sqrt{\frac{S_T^2}{n_T} + \frac{S_R^2}{n_R}}\right)$ for $\mu_T - \mu_R$ lies within the interval (−δ, δ).

Power function of the equivalence test. In this section, we derive the power function of the statistical test to test the hypotheses in (2). We need to consider determining the proper equivalence margin δ first, which is the critical and challenging step in the equivalence test. In this paper, on the basis of previous studies such as those by Chow [9], Tsong et al. [10], and others, we take the equivalence margin δ as a function of the variability of the reference product with the form of $\delta = f \times \sigma_R$, where f is a constant. The variability σ_R is unavailable to the biosimilar developer and is conventionally estimated by sample SD of the reference product. The multiplier f can be adjusted by the pre-given power $1 - \beta$ and the true underlying mean difference between the proposed biosimilar and reference products. Here, the true underlying mean difference is denoted by $\mu_T - \mu_R = \theta$ and it is also considered as a function of σ_R, i.e., $\mu_T - \mu_R = \theta = \eta \times \sigma_R$, where η is a prespecified tolerable shift. Differences in population mean are expected between biosimilar and reference products, because biosimilar products made from living cells or organisms have a much larger variability than do generic drug products. Thus, the equivalence test allows a mean shift of $\eta \times \sigma_R$ and the target mean difference is $\mu_T - \mu_R = \eta \times \sigma_R$.

Under a parallel design and the hypothesis (2), since the $\frac{v(S_T^2/n_T + S_R^2/n_R)}{\sigma_T^2/n_T + \sigma_R^2/n_R}$ approximately follows a chi-squared distribution with v degrees of freedom based on the Welch–Satterthwaite equation [27], the exact power function can be derived by modifying the power formula for the crossover bioequivalence study proposed by Shen et al. [28]:

$$\begin{aligned}
&\text{power} = P(\theta, n_T, n_R, \sigma_T^2, \sigma_R^2) \\
&= P\{\text{Reject } H_0 | \theta = \mu_T - \mu_R, \theta \in (-\delta, \delta), \sigma_T^2, \sigma_R^2\} \\
&= P\left\{\frac{-\delta - \theta}{\sqrt{S_T^2/n_T + S_R^2/n_R}} + t_{\alpha,v} < \frac{(\bar{X}_T - \bar{X}_R) - \theta}{\sqrt{S_T^2/n_T + S_R^2/n_R}} < \frac{\delta - \theta}{\sqrt{S_T^2/n_T + S_R^2/n_R}} - t_{\alpha,v} \middle| \theta \in (-\delta, \delta), \sigma_T^2, \sigma_R^2\right\} \quad (7) \\
&= \int_0^A \left\{\Phi\left(\frac{\delta - \theta}{\sqrt{\sigma_T^2/n_T + \sigma_R^2/n_R}} - t_{\alpha,v}\sqrt{\frac{x}{v}}\right) - \Phi\left(\frac{-\delta - \theta}{\sqrt{\sigma_T^2/n_T + \sigma_R^2/n_R}} + t_{\alpha,v}\sqrt{\frac{x}{v}}\right)\right\} \cdot f(x)\mathrm{d}x,
\end{aligned}$$

where Φ(·) is the standard normal cumulative distribution function and $f(x)$, the probability density function of the chi-squared distribution, can be written as $f(x) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}$. The upper limit of the integral is defined as $A = \frac{v \cdot \delta^2}{t_{\alpha,v}^2 \cdot (\sigma_T^2/n_T + \sigma_R^2/n_R)}$. Formula (7) can be adapted for the equivalence test with equal and unequal variance. We can calculate power values and determine the sample size for the equivalence test in analytical similarity assessment from (7) by using a standard numerical integration. It should be noted that the sample size formula in analytical studies for similarity assessment proposed by Chow et al. [15] is given by $n_T = \frac{(z_\alpha + z_{\beta/2})^2 \sigma^2 (1 + 1/k)}{(\delta - |\mu_T - \mu_R|)^2}$ assuming that $\sigma_R^2 = \sigma_T^2 = \sigma^2$, where $k = n_T/n_R$ and z_α is the upper α quantile of the standard normal distribution (for example, $z_{0.05} = 1.645$). The sample size formula by Chow et al. should be obtained based on the approximate power:

$$\Phi\left(\frac{\delta - \theta}{\sigma\sqrt{1/n_T + 1/n_R}} - z_\alpha\right) + \Phi\left(\frac{\delta + \theta}{\sigma\sqrt{1/n_T + 1/n_R}} - z_\alpha\right) - 1 \approx 2\Phi\left(\frac{\delta - \theta}{\sigma\sqrt{1/n_T + 1/n_R}} - z_\alpha\right) - 1. \quad (8)$$

The above approximate power formula (8) works very well when the sample size is large. It may underestimate the power if the sample size is too small. Therefore, we prefer the explicit formula (7) for sample size determination and various simulation studies.

Using formula (7), we conducted several simulation studies under various parameter settings, including different *f* and η, sample sizes (n_T, n_R), and ratios of variances σ_R^2/σ_T^2. The simulation of various parameter settings is necessary. For example, we may need to increase the constant *f* when sample reference variability may be underestimated if reference values are correlated because of the same source. Under the assumption that $\sigma_T^2 = \sigma_R^2$, S1 and S2 Files provide details of simulation results. S1 File lists the assigned power for different values of the multiplier *f* (from 1 to 2.5 by 0.02) and the given number of lots per product *n* (from 3 to 25 by 1) with $\mu_T - \mu_R = 1/8 \times \sigma_R$ and $\alpha = 0.05$. S2 File gives results of the assigned power for cases of different η values (from 1/16 to 1/2 by 1/16) and the given number of lots per product *n* (from 3 to 25 by 1) with $f = 1.5$ and $\alpha = 0.05$. Note that when we choose the equivalence margin as δ = 1.5 × σ_R and the true mean difference as $\mu_T - \mu_R = 1/8 \times \sigma_R$, $n_T = n_R = 9$ are required to achieve an 80% power at the 5% level of significance. That is, 9 biosimilar and reference lots are sufficient to make meaningful comparisons. Furthermore, the test has 87% power to reject the null hypothesis in favor of equivalence when $n_T = n_R = 10$ with equal variance.

Sample size requirement. Another commonly encountered question is how to handle large sample size imbalance in determining the number of reference lots and the number of test lots required in the equivalence test. As is often the case, the available reference lots denoted by N_R are usually larger than the available biosimilar lots denoted by N_T, because biosimilar developers need a sufficient number of reference lots to understand the reference product. Directly choosing $n_T = N_T$ and $n_R = N_R$ in the above equivalence test may lead to concerns that the information of the reference product can potentially dominate the power of the equivalence test [16]. We can conduct a simulation study to compare power to explain why sample size imbalance needs to be adjusted using formula (7). In Fig 2, we give an example for simulation results for $n_T = 10$. For each n_T, n_R increases from $n_R = n_T$ to $n_R = 5n_T$ and 3 ratios of variances σ_R/σ_T are chosen: 1, $\sqrt{1.5}$, and $\sqrt{2}$. The multiplier η in the true mean difference between the biosimilar and reference products, $\mu_T - \mu_R = \eta \times \sigma_R$, increases from 0 to 1. Fig 2 shows that a biosimilar product with a larger mean difference $\mu_T - \mu_R$ can achieve the desired power by increasing the sample size of 1 arm n_R only. For example, when $\sigma_R/\sigma_T = 1$, $\eta = 8/16$, and $n_T = n_R = 10$, we can increase the power of the equivalence test from about 70% to above 80% by only increasing n_R to 50. To avoid the case in which a large mean difference may be overlooked, we need to adjust sample size imbalance to make $n_T \leq n_R \leq 1.5n_T$.

Chow et al. [15] also proposed that sample size imbalance can be adjusted by the appropriate λ in the relationship $n_R = \lambda \times n_T$. However, both the reference and test lots are often very limited and the coefficient λ is often a decimal and difficult to determine. Thus, we establish a more flexible relationship between the n_R and n_T required as $n_R = n_T + k$ in the equivalence test, where $k = 0,1,\ldots,\lceil 0.5n_T \rceil$; the symbol "$\lceil\ \rceil$" returns the value of a number rounded upward to the nearest integer. The proposed relationship can guarantee that n_R is within $[n_T, 1.5n_T]$ and nearly balanced with n_T, even for sample sizes as small as 10. On the basis of the above relationship and the power function presented in formula (7), we can calculate the minimum n_T for various selections of *k* in the simulation study. Once the mininum n_T has been determined, we can determine the values of *k* and n_R required in the equivalence test.

Table 1 gives examples of simulation results for $1 - \beta$ = 80%, 85%, and 90% when $f = 1.5$ (equivalence margin δ = 1.5 × σ_R) and $\eta = 1/8$ (true underlying mean difference $\mu_T - \mu_R = 1/8 \times \sigma_R$) with $\sigma_R = \sigma_T$. From Table 1, it is easy to determine that the minimum $n_T = 8$ and choose $k = 2$ to satisfy the relationship $n_R \in [n_T, 1.5n_T]$, that is, $(n_T, n_R) = (8,10)$ to achieve an 80%

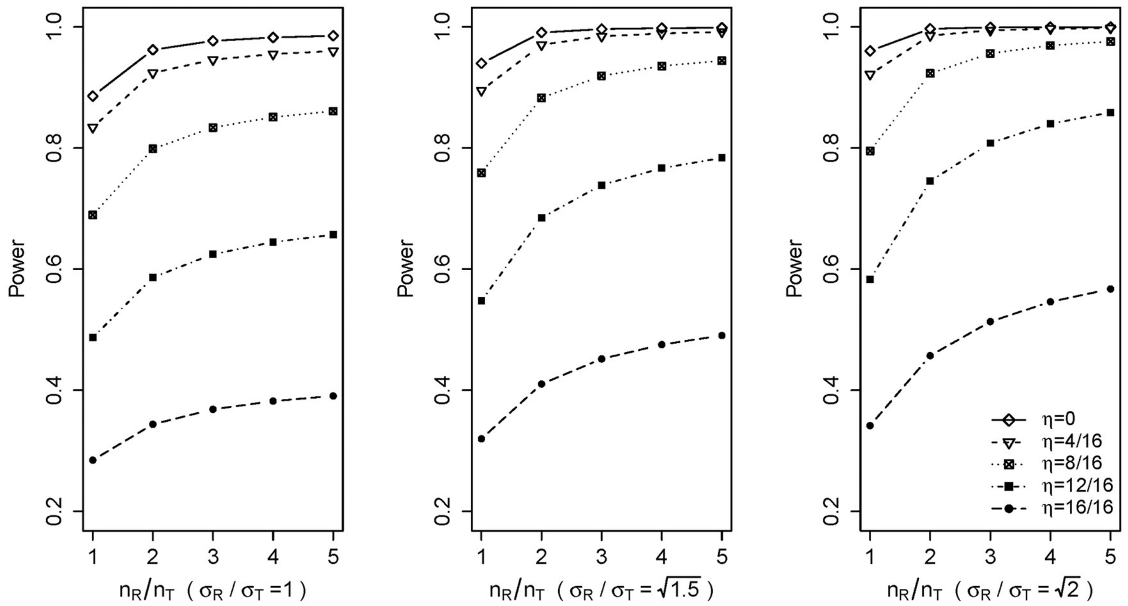

Fig 2. Power with n_T = 10 and margin δ = 1.5 × σ_R at different values of the sample size ratio, variance ratio, and true mean difference.

https://doi.org/10.1371/journal.pone.0208354.g002

power at the 5% level of significance in an equivalence test for CQAs from Tier 1. The combinations $(n_T, n_R) = (7,11),(7,12)$ do not meet the criterion of n_R being within $[n_T, 1.5n_T]$. Obviously, there are many other alternative combinations of sample sizes, such as $(n_T, n_R) = (9,9)$, (9,10), and (8,11). The reason for taking $(n_T, n_R) = (8,10)$ as the optimum combination is that it can ensure the lowest number of sample sizes for biosimilar products. Similarly, the optimum combination is (8,12) for a nominal 85% power and (10,12) for a nominal 90% power.

For different f (from 1 to 2.5 by 0.02) and η (from 1/16 to 1/2 by 1/16) values, the optimum combination (n_T, n_R) with $\alpha = 0.05$, $1 - \beta = 80\%$, 85%, and 90% under the assumption of $\sigma_R = \sigma_T$ are shown in S3 File. Hence, the optimum combinations given first minimize the number of biosimilar lots and then determine n_R based on an appropriate k. Note that if there are enough biosimilar lots N_T, equal sample sizes are preferred to assess analytical similarity, such as $(n_T, n_R) = (9,9)$, (10,10), and (11,11) for achieving $1 - \beta = 80\%$, 85%, and 90%, respectively.

After n_R has been determined on the basis of the above simulation result, n_R needs to be randomly selected from the available reference lots N_R. When selecting n_R from N_R, to reduce the sampling error associated with simple random samples, different n_R lots should be chosen through simulation studies with at least 100,000 replications to determine whether a high proportion (e.g., >80% of these replications) yields the same results in the equivalence test. In

Table 1. Sample size n_T for different k values and powers for $\alpha = 0.05$ with $\sigma_R = \sigma_T$.

Power(1-β)	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
80%	9	9	8	8	7[a]	7[a]
85%	10	10	9	9	8	8[a]
90%	11	11	10	10	10	9[a]

[a] Value does not meet the criterion that n_R is within$[n_T, 1.5n_T]$.

https://doi.org/10.1371/journal.pone.0208354.t001

practice, we use the entire available reference lots N_R to estimate σ_R to establish the equivalence margin $\delta = f \times \sigma_R$.

Mann–Whitney test for equivalence

The above discussion demonstrates that the sample size in the equivalence test for CQAs from Tier 1 is relatively small. In this situation, the assumption of normality for data may be violated, and a distribution-free or nonparametric test may be more appropriate for comparing these independent samples. We consider using the Mann–Whitney test for equivalence, which is sensitive to divergences between any 2 continuous distributions. For simplicity, let T_i and R_j be observations of the biosimilar and reference arms. If the 2 distributions of T_i and R_j are equivalent, then the probability that any value of T_i is greater than any value of R_j denoted by $\pi_+ = P[T_i > R_j]$ should be approximately 1/2. Alternatively, the null hypothesis is that π_+ is either smaller or larger than the range of equivalence. Therefore, the Mann–Whitney test for equivalence uses a rank-sum statistic to test whether π_+ is within the small range of approximately 1/2. Thus, the equivalence hypothesis for the non-parametric test of testing problem (2) is given by

$$H_0 : \pi_+ \leq 1/2 - \delta' \text{ or } \pi_+ \geq 1/2 + \delta' \text{ vs. } H_a : 1/2 - \delta' < \pi_+ < 1/2 + \delta', \tag{9}$$

where δ' is defined by $\delta' = \Phi(\delta/\sqrt{2\sigma^2}) - 1/2$, where σ is the pooled standard deviation of T_i and R_j. The value π_+ is estimated using the Mann–Whitney statistic, and the estimator W_+ defined as $W_+ = 1/(n_T n_R)\sum_{i=1}^{n_T}\sum_{j=1}^{n_R} I(T_i - R_j)$ is given with the indicator of a positive sign

$$I(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}.$$

Rejecting the nonequivalence H_0 in (9) if and only if

$$|W_+ - 1/2|/\hat{\sigma}_{W_+} < C(\alpha, \delta'), \tag{10}$$

where

$$\hat{\sigma}^2_{W_+} = \frac{1}{n_T n_R}(W_+ - (n_T + n_R - 1)\, W_+^2 + (n_T - 1)\prod\nolimits_{XXY} + (n_R - 1)\prod\nolimits_{XYY}),$$

$$\prod\nolimits_{XXY} = \frac{2}{n_T n_R (n_T - 1)} \sum_{i_1=1}^{n_T-1} \sum_{i_2=i_1+1}^{n_T} \sum_{j=1}^{n_R} I(T_{i_1} - R_j) \cdot I(T_{i_2} - R_j),$$

and

$$\prod\nolimits_{XYY} = \frac{2}{n_T n_R (n_R - 1)} \sum_{i=1}^{n_T} \sum_{j_1=1}^{n_R-1} \sum_{j_2=j_1+1}^{n_R} I(T_i - R_{j_1}) \cdot I(T_i - R_{j_2}),$$

and $C^2(\alpha, \delta')$ is the α 100%th percentile of the non-central chi-squared distribution with degrees of freedom equal to 1 and positive noncentrality parameters equal to $\delta'^2/\hat{\sigma}^2_{W_+}$. The Mann–Whitney test for equivalence is asymptotically distribution free with respect to the significance level and controls the level even for sample sizes as small as 10. Details of the derivation process of formulas and the calculation method have been rigorously established by Wellek [29].

So far, we have developed an analytical similarity evaluation roadmap that includes our proposed statistical approaches for CQAs from Tier 1. Key steps of the roadmap are described as follows:

Step 1: Determine the CQAs from Tier 1 through the systematic risk ranking and tiering approach we introduced.

Step 2: Determine the margin as given in S1–S3 Files, select n_T, k, n_R, and then determine the sample size.

Step 3: Conduct the equivalence test or Mann–Whitney test for equivalence for CQAs of interest from Tier 1 and draw relevant conclusions.

Case study

In this case study, we have acquired the analytical data for 2 CQAs from a pharmaceutical company, to show how our proposed statistical evaluation roadmap can be used to assess analytical similarity. Because of the commercial confidentiality, sensitive information such as the name of the CQA is masked and data are used only as examples to validate the methods for both equivalence test and Mann-Whitney test.

The 2 CQAs have been identified by relevant company, especially researchers in the quality control team, and based on the risk ranking and tier assignment approach that we previously introduced. Numerical values are assigned to impact and uncertainty and multiplied to generate a relative risk score. Finally, the 2 CQAs having the highest risk ranking among attributes and are suited for statistical tests are considered the most relevant to clinical outcomes assigned to Tier 1 after a rigorous internal discussion among drug developers. S4 File gives analytical data for CQA1 and CQA2 of the reference and test groups. Analytical data include 11 lots of the test group and 61 lots of the reference group for CQA1, and 11 lots of the test group and 50 lots of the reference group for CQA2. Analytical data of 2 CQAs from each lot are shown in Figs 3 and 4, respectively. Both figures show large overlaps between the test and reference groups. It is clear that the sample size for the reference group, denoted by N_R, is larger than that for the test group, denoted by N_T, that is, $N_R \gg N_T$. Table 2 shows summary statistics for the 2 CQAs.

Using CQA1 as an example, we can perform a similar analysis for CQA2. Table 3 summarizes the parameter settings and results of statistical evaluation. First, CQA1 undergoes the statistical equivalence test. To compare the reference and test groups, sufficient communication is needed with drug developers. Then, the multiplier $f = 1.5$ for the margin $\delta = f \times \sigma_R$ and the multiplier $\eta = 1/8$ for the true underlying mean difference $\mu_T - \mu_R = 1/8 \times \sigma_R$ is determined. Since N_R is much larger than N_T in CQA1, it is not appropriate to directly make $n_T = N_T$ and $n_R = N_R$ in the equivalence test and it is necessary to make some adjustments for imbalanced sample size. We first determine that $n_T = N_T = 11$ and then divide the reference lots N_R into 2 parts according the n_R required. As shown in S1 File, under $\delta = 1.5 \times \sigma_R$ and $\mu_T - \mu_R = 1/8 \times \sigma_R$, the power achieved is nearly 91% at the 5% level of significance when the sample size is 11 for both the groups. Hence, we choose $k = 0$ and make $n_R = n_T + k = 11$. To establish the equivalence margin δ, we use the entire available reference lots N_R to estimate σ_R. Consequently, we obtain $(n_T, n_R) = (11, 11)$ and margin = (−1.17,1.17) in the equivalence test for CQA1. As shown in Table 3, the high proportion (97.66%) of CI of 10^5 random samples is completely within the margin (−1.17,1.17) for CQA1. Here, we also list results of the Mann–Whitney test for equivalence with $(n_T, n_R) = (11, 11)$ and margin = (0.13,0.87). The Mann–Whitney test could lose power when the normality assumption for data is valid. In this case study, we

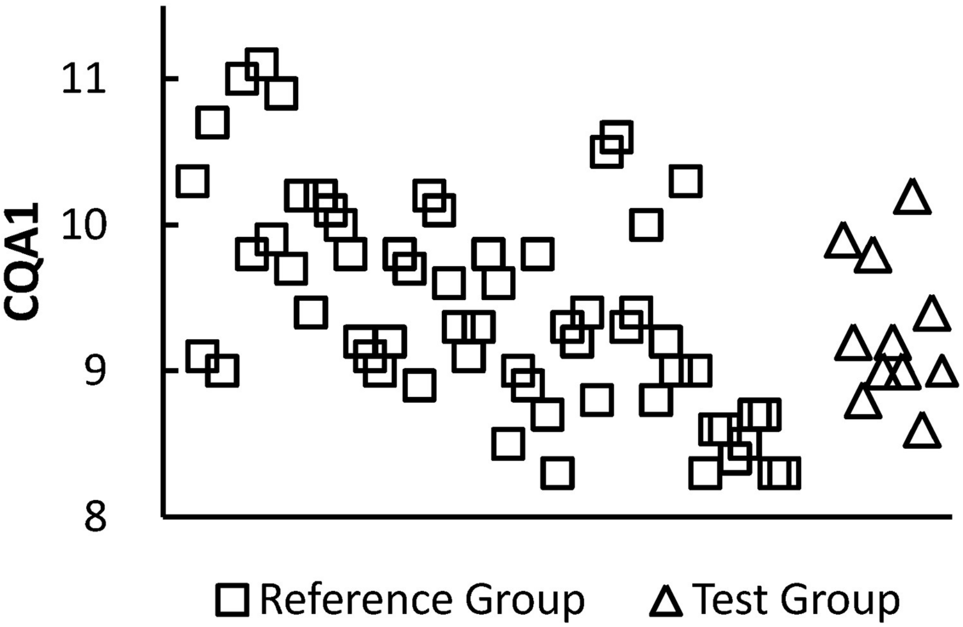

Fig 3. Analytical data for CQA1 from each lot. CQA: critical quality attribute.

https://doi.org/10.1371/journal.pone.0208354.g003

claim that the CQA1 of 2 groups is analytically similar, based on results of the equivalence test, because the analytical data are approximately normally distributed. If the analytical data have a seriously skewed distribution, we will make a decision based on results of the Mann–Whitney test.

In summary, statistical evaluations for the 2 CQAs demonstrate the analytical similarity between the reference and test groups. R programs are provided in S5 File for readers to get

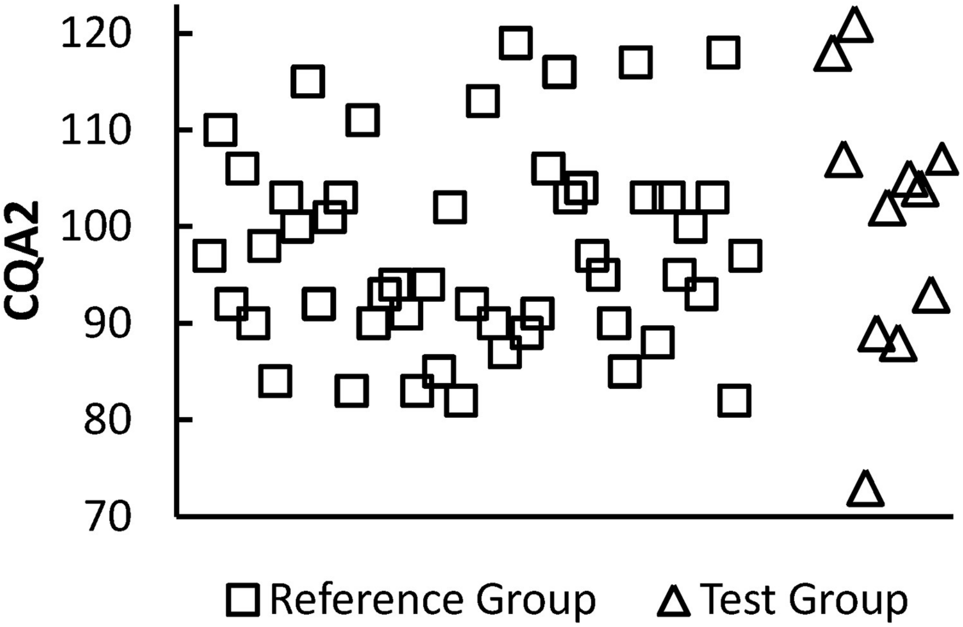

Fig 4. Analytical data for CQA2 from each lot. CQA: critical quality attribute.

https://doi.org/10.1371/journal.pone.0208354.g004

Table 2. Summary statistics for CQA1 and CQA2.

Statistics	CQA1		CQA2	
	RG	TG	RG	TG
Number of lots	61	11	50	11
Mean	9.46	9.28	97.50	100.64
SD	0.78	0.50	10.15	13.95
%CV	8.28	5.34	10.41	13.86
P-value[a]	0.049	0.428	0.048	0.705

CQA: critical quality attribute; RG: reference group; TG: test group; SD: standard deviation; CV: coefficient of variation.

[a] *P*-values were calculated for the Shapiro–Wilk normality test.

https://doi.org/10.1371/journal.pone.0208354.t002

detailed results using the proposed methods, including the equivalence test and the Mann–Whitney test for equivalence.

Conclusions

We propose a statistical evaluation roadmap using feasible statistical methods for analytical similarity assessment of CQAs from Tier 1. The statistical evaluation roadmap has 3 advantages: (i) there is a very flexible relationship between n_R and n_T, as $n_R = n_T + k$ in the equivalence test; (ii) there is much more flexibility in choosing parameters such as equivalence margins and the true underlying mean difference as well as in obtaining optimum sample sizes; and (iii) the Mann–Whitney test is used for analytical data that follow a skewed distribution. Using this roadmap, we found sufficiently strong evidence to support the similarity between the reference and biosimilar products. A sufficient degree of biosimilarity demonstrated in the earlier step of head-to-head analytical assessment can serve as a foundation to develop biosimilars and facilitate an abbreviated subsequent preclinical and clinical evaluation, thus enabling a shorter path to licensing. This is different from the typical development of a new small-molecule drug, wherein the pathway heavily focuses on the endpoints of clinical evaluations relating to demonstrating efficacy and safety in humans.

Although there are several advantages of the proposed roadmap, there are still some unsolved issues. First, the variability of the reference is underestimated when the method does

Table 3. Summarized results of statistical evaluation for CQA1 and CQA2.

Test of conduct	Parameter	CQA1	CQA2
Equivalence test of means	Equivalence margin [a]	(–1.17,1.17)	(–15.23,15.23)
	Sample sizes (n_T, n_R)	(11, 11)	(11, 11)
	Random samples	10^5	10^5
	Proportion	97.66%	88.83%
Conclusion: Analytically similar		Yes	Yes
Mann–Whitney test for equivalence	Equivalence margin [b]	(0.13,0.87)	(0.16,0.84)
	Sample sizes (n_T, n_R)	(11, 11)	(11, 11)
	Random samples	10^5	10^5
	Proportion	93.60%	73.19%

CQA, critical quality attribute.

[a] Margin of the equivalence test is $(-1.5\hat{\sigma}_R, 1.5\hat{\sigma}_R)$.

[b] Margin of the Mann–Whitney test is $(1-\Phi(1.5\hat{\sigma}_R/\sqrt{2\sigma^2}), \Phi(1.5\hat{\sigma}_R/\sqrt{2\sigma^2}))$.

https://doi.org/10.1371/journal.pone.0208354.t003

not consider the case in which we sample more than one item from each lot, which leads to a conservative test and affects sample size determination [30]. Second, when the available reference lots N_R are larger than the available biosimilar lots N_T, the n_R lots required in the equivalence test need to be randomly selected from N_R. Thus, the N_R lots are divided into 2 parts: n_R and $N_R - n_R$. We use the entire data of N_R lots to estimate σ_R to establish the equivalence margin in our evaluation roadmap. Further discussion is required for the case in which the first part contains the n_R lots or the second part contains the remaining reference sample $N_R - n_R$ lots used to determine the equivalence margin. Our future studies will focus on incorporating these challenges into the current proposed framework.

Supporting information

S1 File. Assigned power for *f* and *n*.
(XLS)

S2 File. Assigned power for *η* and *n*.
(XLS)

S3 File. Optimum combination of sample size.
(XLSX)

S4 File. Analytical testing value of CQA1 and CQA2.
(XLSX)

S5 File. R programs.
(DOC)

Acknowledgments

The authors thank Vani Shanker, PhD, Department of Scientific Editing, St. Jude Children's Research Hospital, for editing the manuscript.

Author Contributions

Conceptualization: Haitao Pan, Jielai Xia.

Data curation: Kejian Wu, Haitao Pan.

Formal analysis: Kejian Wu, Haitao Pan.

Funding acquisition: Ling Wang, Jielai Xia.

Investigation: Kejian Wu, Ling Wang.

Methodology: Kejian Wu, Haitao Pan, Jielai Xia.

Project administration: Qingbo Zhao, Jielai Xia.

Resources: Chen Li, Qingbo Zhao.

Software: Kejian Wu, Haitao Pan.

Supervision: Qingbo Zhao, Ling Wang, Jielai Xia.

Validation: Kejian Wu, Chen Li.

Visualization: Kejian Wu, Chen Li.

Writing – original draft: Kejian Wu, Haitao Pan.

Writing – review & editing: Kejian Wu, Haitao Pan, Ling Wang, Jielai Xia.

References

1. Stevenson JG, Popovian R, Jacobs I, Hurst S, Shane LG. Biosimilars: Practical considerations for pharmacists. Ann Pharmacother. 2017; 51(7): 590–602. https://doi.org/10.1177/1060028017690743 PMID: 28176529
2. Generics and Biosimilars Initiative (GaBi). FDA approves epoetin alfa biosimilar Retacrit. http://www.gabionline.net/Biosimilars/News/FDA-approves-epoetin-alfa-biosimilar-Retacrit.
3. Udpa N, Million RP. Monoclonal antibody biosimilars. Nature reviews. Nat Rev Drug Discov. 2016; 15, 13–14. https://doi.org/10.1038/nrd.2015.12 PMID: 26678619
4. Guideline on similar biological medicinal products. European Medicines Agency. 2015. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/10/WC500176768.pdf
5. Scientific considerations in demonstrating biosimilarity to a reference product. Food and Drug Administration. 2015. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf
6. Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product. Guidance for industry. Food and Drug Administration. 2015. https://www.fda.gov/downloads/drugs/guidances/ucm291134.pdf
7. World Health Organization. Guidelines on evaluation of similar biotherapeutic products (SBPs). Geneva: World Health Organization. 2009. http://www.who.int/biologicals/publications/trs/areas/biological_therapeutics/TRS_977_Annex_2.pdf
8. China Food and Drug Administration. Draft guideline on development and evaluation of biosimilars (Chinese Version). 2015. http://samr.cfda.gov.cn/WS01/CL1616/115104.html
9. Chow SC. On assessment of analytical similarity in biosimilar studies. Drug Des. 2014; 3, 2138–2169. https://doi.org/10.4172/2169-0138.1000e124
10. Tsong Y, Dong X, Shen M. Development of statistical methods for analytical similarity assessment. J Biopharm Stat. 2017; 27, 197–205. https://doi.org/10.1080/10543406.2016.1272606 PMID: 27977326
11. Liu J, Eris T, Li C, Cao S, Kuhns S. Assessing analytical similarity of proposed amgen biosimilar ABP 501 to adalimumab. BioDrugs. 2016; 30, 321–338. https://doi.org/10.1007/s40259-016-0184-3 PMID: 27461107
12. Seo N, Polozova A, Zhang M, Yates Z, Cao S, Li H, et al. Analytical and functional similarity of Amgen biosimilar ABP 215 to bevacizumab. Mabs. 2018; 10, 678–691. https://doi.org/10.1080/19420862.2018.1452580 PMID: 29553864
13. Chow SC, Song F, Bai H. Analytical similarity assessment in biosimilar studies. AAPS J. 2016; 18, 670–677. https://doi.org/10.1208/s12248-016-9882-5 PMID: 26873509
14. Chow SC. Challenging issues in assessing analytical similarity in biosimilar studies. Biosimilars. 2015; 33–39. https://doi.org/10.2147/BS.S84141
15. Chow SC, Song F, Bai H. Sample size requirement in analytical studies for similarity assessment. J Biopharm Stat. 2017; 27, 233–238. https://doi.org/10.1080/10543406.2016.1265545 PMID: 27935446
16. Dong XC, Weng YT, Tsong Y. Adjustment for unbalanced sample size for analytical biosimilar equivalence assessment. J Biopharm Stat. 2017; 27, 220–232. https://doi.org/10.1080/10543406.2016.1265544 PMID: 28060570
17. Shen M, Wang T, Tsong Y. Statistical considerations regarding correlated lots in analytical biosimilar equivalence test. J Biopharm Stat. 2017; 27, 213–219. https://doi.org/10.1080/10543406.2016.1265541 PMID: 27906604
18. Burdick R, Coffey T, Gutka H, Gratzl G, Conlon HD, Huang CT, et al. Statistical approaches to assess biosimilarity from analytical data. AAPS J. 2017; 19, 4–14. https://doi.org/10.1208/s12248-016-9968-0 PMID: 27709452
19. Dong XC, Bian Y, Tsong Y, Wang T. Exact test-based approach for equivalence test with parameter margin. J Biopharm Stat. 2017; 27, 317–330. https://doi.org/10.1080/10543406.2016.1265546 PMID: 28055327
20. Chen YM, Weng YT, Dong X, Tsong Y. Wald tests for variance-adjusted equivalence assessment with normal endpoints. J Biopharm Stat. 2017; 27, 308–316. https://doi.org/10.1080/10543406.2016.1265542 PMID: 27906607
21. Liao JJ, Darken PF. Comparability of critical quality attributes for establishing biosimilarity. Stat Med. 2013; 32, 462–469. https://doi.org/10.1002/sim.5564 PMID: 22903263

22. Wu KJ, Pan HT, Zhao QB, Li CJ, L C, W L, et al. Some statistical considerations in analytical similarity assessment of biosimilar studies. Chinese Journal of Health Statistics. 2018; 35, 343–348.

23. Alt N, Zhang TY, Motchnik P, Taticek R, Quarmby V, Schlothauer T, et al. Determination of critical quality attributes for monoclonal antibodies using quality by design principles. Biologicals. 2016; 44, 291–305. https://doi.org/10.1016/j.biologicals.2016.06.005 PMID: 27461239

24. Vandekerckhove K, Seidl A, Gutka H, Kumar M, Gratzl G, Keire D, et al. Rational selection, criticality assessment, and tiering of quality attributes and test methods for analytical similarity evaluation of biosimilars. AAPS J. 2018; 20, 68. https://doi.org/10.1208/s12248-018-0230-9 PMID: 29748754

25. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1983; 15, 657–680. https://doi.org/10.1007/bf01068419

26. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics Bulletin. 1946; 2, 110–114. https://doi.org/10.2307/3002019 PMID: 20287815

27. Welch BL. The generalization of 'Student's' problem when several different population variances are involved. Biometrika. 1947; 34(1/2): 28–35. https://doi.org/10.2307/2332510

28. Shen M, Russek-Cohen E, Slud EV. Exact calculation of power and sample size in bioequivalence studies using two one-sided tests. Pharm Stat. 2015; 14(2): 95–101. https://doi.org/10.1002/pst.1666 PMID: 25477145

29. Wellek S. A new approach to equivalence assessment in standard comparative bioavailability trials by means of the Mann-Whitney statistic. Biom J. 1996; 38, 695–710. https://doi.org/10.1002/bimj.4710380608

30. Wang T, Chow SC. On the establishment of equivalence acceptance criterion in analytical similarity assessment. J Biopharm Stat. 2017; 27, 206–212. https://doi.org/10.1080/10543406.2016.1265539 PMID: 28051920