

Establishment of reference standards in biosimilar studies

Aijing Zhang, PhD; Associate Professor Jung-Ying Tzeng; Professor Shein-Chung Chow

When an innovative biological product goes off-patent, biopharmaceutical or biotechnological companies may file an application for regulatory approval of biosimilar products. In practice, however, important information on the innovative (reference) product may not be available for assessment. Thus, it is important to first establish a reference standard while assessing biosimilarity between a biosimilar product and the reference product. In this paper, reference standard is established through the biosimilarity index approach based on a reference-replicated study (or R-R study), in which the reference product is compared with itself under various scenarios. The reference standard can then be used for assessing the degree of similarity between the test and reference drugs in biosimilar studies.

Keywords: Biosimilarity, biosimilarity index, highly similar, reference standards, reference-replicated study

Introduction

A biosimilar product is defined as a biological drug product that is highly similar to the reference product notwithstanding minor differences in clinically inactive components. No clinically meaningful differences exist in terms of safety, purity, and potency (Biologics Price Competition and Innovation [BPCI] Act 2009). In the BPCI Act, however, how 'similar' is considered 'highly similar' has not been addressed. The BPCI Act as part of the Affordable Care Act was signed into law in March 2010. It gave FDA the authority to approve similar biological drug products. Currently, FDA has not set out clear standards for biosimilar approvals [1].

Kang and Chow [2] proposed a method that uses the relative difference, i.e. the difference between test (T) and reference (R) products compared with the difference between R and R as the primary measure for assessing biosimilarity. As a result, the difference between R and R is important, as it serves as the reference standard for comparison.

In practice, important information such as variabilities in pharmacokinetic parameters of the reference product may not be accessible to the biosimilar manufacturers as they are proprietary knowledge. Thus, it is important to conduct a reference-replicated study (R-R study), i.e. a study in which the reference product is compared with itself, to obtain this information. As biosimilar products are made by living cells or organisms that are sensitive to small changes or variations in environmental factors, such as light or temperature, it is expected that a small difference will occur between two reference products resulting from different manufacturing processes or locations, or different batches from the same manufacturing process. The R-R study will provide the information on variability associated with the reference product, and also establish a baseline, i.e. similarity between R and R, for comparison in biosimilarity testing.

The biosimilarity index approach was proposed by Chow et al. [3, 4] to assess the level of similarity. The index is based on the concept of reproducibility probability and is evaluated as the estimated power of the testing procedure when the alternative hypothesis is true. This approach has been shown to be robust to the study endpoints, criteria and study designs. Thus, in this

paper, we have chosen to establish the reference standards in the framework of this biosimilarity index approach.

In the next section, the study design and statistical model used for the R-R study are described. In Section 3, biosimilarity index based on reproducibility probability is introduced. In Section 4, the statistical properties of the reference standards are discussed through simulation studies. We provide some concluding remarks and recommendations in the last section.

Study design and statistical model

Study design

Crossover designs are often recommended for bioequivalence studies. Biosimilarity is often assessed in a parallel design in practice [5], however, because of the relatively long half-life and the possibility of immunogenicity of biological products, concerns are increasing over interchangeability in biosimilars, which can only be assessed in crossover designs. Thus, we have chosen a standard two-sequence, two-period (2×2) crossover design for the establishment of reference standards. A 2×2 crossover study can be viewed as a two-arm parallel study with replicates; the advantage of using a crossover design is that the intra-subject variability could be estimated for the reference standards.

In a 2×2 standard crossover study design, participants are randomly assigned to either sequence 1 (R_1R_2) or sequence 2 (R_2R_1) in the R-R study. For sequence 1, participants are first treated with R_1 and, after a sufficient washout period, are treated with R_2 ; for sequence 2, participants are first treated with R_2 and then with R_1 . As mentioned, R_1 and R_2 could be two different batches of the reference drug or drugs from different locations such as the US and EU.

Statistical model

The interval hypotheses of the similarity testing written in terms of parameters are:

$$\begin{aligned} H_0 &: \theta \leq \theta_L \text{ or } \theta \geq \theta_U \\ H_a &: \theta < \theta < \theta_U \end{aligned} \quad (1)$$

where θ is the study parameter chosen to assess biosimilarity; θ_L and θ_U are the biosimilarity limits, i.e. the accepted lower

Author for correspondence: Aijing Zhang, PhD, 5109 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695, USA

Submitted: 24 April 2013; Revised: 10 June 2013; Edited: 8 July 2013; Accepted: 16 July 2013; Published online first: 31 July 2013

ORIGINAL RESEARCH

Biosimilarity and Interchangeability

and upper bounds for declaring biosimilarity. In bioequivalence assessment for generic drug products, average bioequivalence criterion chooses θ as the ratio of the geometric means between reference and test products; and the corresponding limits are set to be 80% and 125%.

The hypotheses in equation (1) are usually decomposed into two sets of one-sided hypotheses:

$$\begin{aligned} H_{\theta 1} : \theta \leq \theta_L \text{ and } H_{\theta 2} : \theta > \theta_U \\ H_{\alpha 1} : \theta > \theta_L \quad H_{\alpha 2} : \theta > \theta_U \end{aligned} \quad (2)$$

Schuirmann's two one-sided tests procedure is commonly used for testing the hypotheses in equation (2) [6]. The overall p value is the larger of the p values from these two tests.

For a standard 2 x 2 crossover study design, we consider the following statistical model with the assumption of no carryover effects:

$$Y_{ijk} = \mu + S_{ik} + P_j + T_{(j,k)} + \epsilon_{ijk} \quad (3)$$

where Y_{ijk} is the response for subject i in the k th sequence at the j th period, where $i = 1, \dots, n_k$ indicates subject, $j = 1, 2$ indicates period, $k = 1, 2$ indicates sequence; μ represents the overall mean; S_{ik} represents the random effect of i th subject in k th sequence, assumed independently and identically distributed (i.i.d.) as $N(0, \sigma_{s2})$; P_j is the period effect; $T_{(j,k)}$ represents the fixed effect of the treatment in the k th sequence administered at the j th period; ϵ_{ijk} is the within-subject random error, assumed i.i.d. as $N(0, \sigma_{\epsilon 2})$. Finally S_{ik} 's and ϵ_{ijk} 's are assumed to be mutually independent.

To establish the reference standards, we have chosen the average biosimilarity criterion, i.e. $\theta = \mu_{R2} - \mu_{R1}$. One concludes average biosimilarity if

$$T_L = \frac{(\bar{Y}_{R2} - \bar{Y}_{R1}) - \theta_L}{\sigma_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t(\alpha, n_1 + n_2 - 2) \quad (4)$$

and

$$T_U = \frac{(\bar{Y}_{R2} - \bar{Y}_{R1}) - \theta_U}{\sigma_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t(\alpha, n_1 + n_2 - 2) \quad (5)$$

where \bar{Y}_{R1} and \bar{Y}_{R2} are the least square means for the two reference products; they can be obtained from the sequence-by-period means: $\bar{Y}_{R1} = \frac{1}{2}(\bar{Y}_{11} + \bar{Y}_{22})$ and $\bar{Y}_{R2} = \frac{1}{2}(\bar{Y}_{12} + \bar{Y}_{21})$; n_1 and n_2 are the number of subjects in sequence 1 and sequence 2; σ_d is the pooled sample variance of period difference from both sequences, and can be obtained from $\hat{\sigma}_d = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^2 \sum_{i=1}^{n_k} (d_{ik} - \bar{d}_k)^2$, where $d_{ik} = \frac{1}{2}(Y_{i2k} - Y_{i1k}), i = 1, 2, \dots, n_k; k = 1, 2$.

Biosimilarity index

Definition

To reflect the characteristics and effect of variability on the therapeutic effect of biological products, Chow et al. [3, 4] in 2011 proposed the development of an index based on the concept of the reproducibility probability to assess 'highly biosimilar'.

Reproducibility probability in clinical trials was first proposed in 2002 by Shao and Chow [7] to provide assessment on the strength of evidence against null hypothesis for regulatory agencies in deciding whether a single clinical trial is sufficient for approval. Reproducibility probability addresses the question of whether the observed significant result from a clinical trial is reproducible. Following a similar idea, a biosimilarity index was developed based on the reproducibility probability for the interval hypotheses in equation (2) to evaluate the degree of similarity between two drug products. That is, we define a reproducibility probability for evaluating the biosimilarity of two drug products as follows:

P_{BI} = concluding average biosimilarity between two products in a future trial given that the average biosimilarity based on a chosen criterion has been established in the first trial.

Evaluation

To evaluate the biosimilarity index, the index is expressed as the estimated power of the testing procedure when the alternative hypothesis is true, replacing the parameter by its estimate based on the data observed. By this estimated power approach, the biosimilarity index \hat{P}_{BI} for the 2 x 2 crossover design and average biosimilarity criterion can be obtained from the following equation:

$$\hat{P}_{BI} = P(T_L(Y) > t_L \text{ and } T_U(Y) < t_U | \hat{\delta}_L, \hat{\delta}_U) \quad (6)$$

where T_L and T_U are the test statistics given in equation (4) and equation (5). Both T_L and T_U follow non-central t-distribution, with $n_1 + n_2 - 2$ degree of freedom and non-centrality parameters δ_L and δ_U , respectively. δ_L and δ_U relate to the population means, variances and similarity limits; their estimate $\hat{\delta}_L$ and $\hat{\delta}_U$ can be obtained from the data using the formulas

$$\hat{\delta}_L = \frac{(\bar{Y}_{R2} - \bar{Y}_{R1} - \theta_L)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ and } \hat{\delta}_U = \frac{(\bar{Y}_{R2} - \bar{Y}_{R1} - \theta_U)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Steps

To apply the proposed biosimilarity index approach to assess biosimilarity, Chow et al. in 2013 proposed the following steps [8], see Table 1.

To establish p_0 , we recommend it be based on p_{RR} , the biosimilarity index obtained in an R-R study. By basing p_0 on p_{RR} , the

Table 1: The biosimilarity index approach to assess biosimilarity [8]

Step 1	Assess the average biosimilarity based on a given biosimilarity criterion. The criterion could be based on mean, ratio, or variability.
Step 2	Once the product passes the test for biosimilarity in Step 1, calculate biosimilarity index on the basis of observed mean difference and standard deviation. The calculated biosimilarity index thus takes the variability and the sensitivity of heterogeneity in variances into consideration.
Step 3	We then claim highly biosimilar if the calculated 95% confidence lower bound of the biosimilarity index is larger than p_0 , a pre-specified limit on declaring highly biosimilar.

biosimilarity index approach allows us to assess the degree of similarity relative to the reference product.

Merits

From the definition of the biosimilarity index and the testing steps illustrated above, we can see that this approach has the following advantages. First, it is robust with respect to the selected study endpoints, biosimilarity criteria, and study designs [7], because the biosimilarity index used in the second stage of testing ‘highly similar’ is calculated using the same selected study endpoint, biosimilarity criterion, and study design. Second, it takes variability into consideration in the calculation of the index, and is sensitive to the variance of the test products. Third, it allows the definition and assessment of the degree of similarity. Or, in other words, it provides an answer to the question of ‘how similar is considered highly similar’.

Establishing reference standard

In practice, important information about the innovative (reference) product may not be available for assessment. Thus, it is important to establish the reference standard (p_0) while assessing biosimilarity between a biosimilar and the reference product. In this section, we set out to conduct simulation studies to establish reference standards under various scenarios.

Simulation design

Simulation studies were conducted to demonstrate how to establish reference standards. Standard 2 x 2 crossover design and average biosimilarity criterion are used. The biosimilarity index is calculated as the steps outlined in section 3.

The study parameter θ is the mean difference, i.e. $\theta = \mu_{R2} - \mu_{R1}$, and the corresponding similarity limits are those accepted in bioequivalence studies, i.e. $(\theta_L, \theta_U) = (\log 0.8, \log 1.25) = (-0.223, 0.223)$. The data are assumed to following normal distribution, and the means and variances of the two reference products are assumed to be the same in the simulation, i.e. $\mu_{R1} = \mu_{R2} = \mu$ and $\sigma_{R1} = \sigma_{R2} = \sigma$. To evaluate how reference standards are affected by variability of the drug products, different coefficients of variation are chosen: 20–50% by an increment of 5%. Note that coefficients of variation is the ratio of the standard deviation to the mean, therefore in terms of σ and μ in the normal distribution, the chosen coefficients of variation values translate to $\sigma = 0.2\mu, 0.25\mu, 0.3\mu, 0.35\mu, 0.4\mu, 0.45\mu$ or 0.5μ . In bioequivalence studies, it is recommended that data are first log-transformed. For lognormal distribution, coefficients of variation of 20%, 25%, 30%, 35%, 40%, 45% and 50% translate to $\sigma = 0.2, 0.25, 0.294, 0.340, 0.385, 0.429$ and 0.472 ($\sigma = \sqrt{\ln(1 + CV^2)}$). We have chosen $\mu = 1$ so that the results are comparable with those shown in original scales in bioequivalence studies. The effects of different sample sizes ($n = n_1 + n_2 = 40, 50, 60, 70, 80, 90$ and 100) with equal allocation are also investigated. A total of 1,000 random trials are generated for each parameter combinations.

Results

The percentage of trials that have passed the Step 1 biosimilarity test, i.e. the probability of claiming biosimilarity on the basis of the average biosimilarity criterion, are shown in Table 2 and Figure 1. The results are in accordance with results observed in other studies [9]. As the variability increases, the probability

Table 2: Percentage of trials that passed the Step 1 test

Sample size (n)	Coefficient of variation						
	20%	25%	30%	35%	40%	45%	50%
40	99.9	98.5	91.4	79.2	63.0	46.6	30.5
50	100	99.7	96.8	89.0	75.7	59.3	45.5
60	100	99.9	98.3	93.2	83.7	69.1	55.9
70	100	100	99.5	96.5	90.5	77.8	66.0
80	100	100	99.7	98.2	94.4	85.0	74.9
90	100	100	99.9	99.1	96.8	89.7	81.0
100	100	100	99.9	99.6	97.8	93.6	86.0

decreases. Even with no mean difference between the two drug products, the probability of concluding biosimilarity can be quite low when the coefficients of variation are high. For example, when the coefficients of variation is 50%, with a total sample size of 40, biosimilarity is concluded only about 30% of the times. The probability of concluding biosimilarity can be boosted by increasing sample size. In the previous scenario, the probability can be increased from 30% to 86% if a sample size of 100 is used instead of 40.

For those trials that have passed the Step 1 test, the p values from the Schuirmann’s two one-sided tests procedure are recorded and the averages are reported in Table 3 and Figure 2. As the variability increases, the p values also increase, becoming close

Figure 1: Percentage of trials that passed the Step 1 test

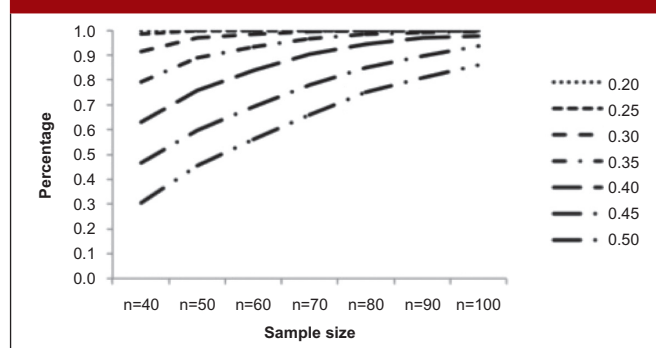
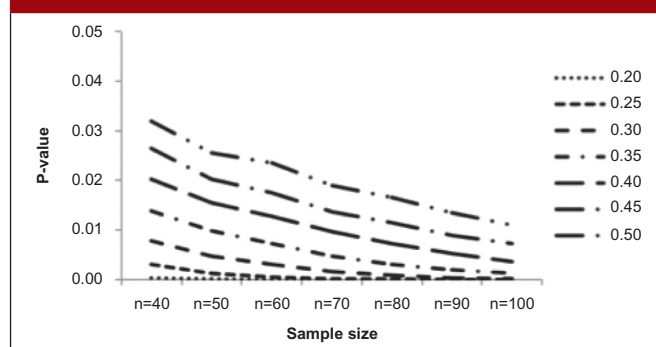


Figure 2: Average p values for trials that passed the Step 1 test



ORIGINAL RESEARCH

Biosimilarity and Interchangeability

Table 3: Average p values for trials that passed the Step 1 test

Sample size (n)	Coefficient of variation						
	20%	25%	30%	35%	40%	45%	50%
40	0.0004	0.0030	0.0078	0.0139	0.0202	0.0265	0.0320
50	0.0001	0.0012	0.0047	0.0099	0.0156	0.0203	0.0256
60	< 0.0001	0.0006	0.0030	0.0072	0.0127	0.0176	0.0236
70	< 0.0001	0.0002	0.0017	0.0048	0.0097	0.0137	0.0189
80	< 0.0001	0.0001	0.0008	0.0031	0.0073	0.0115	0.0167
90	< 0.0001	< 0.0001	0.0004	0.0020	0.0053	0.0089	0.0135
100	< 0.0001	< 0.0001	0.0002	0.0013	0.0036	0.0072	0.0110

to 0.05. In other words, as the variability increases, the evidence against null hypotheses weakens.

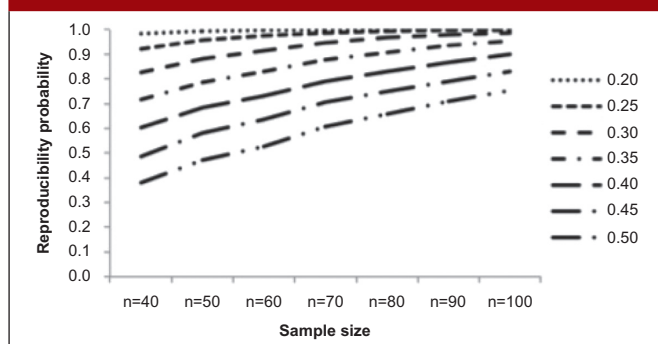
For those trials that have passed the Step 1 test, we further report in Table 4 and Figure 3 the reproducibility probability or biosimilarity index, termed p_{RR} in R-R study. The p_{RR} is calculated from equation (6). As expected, the results show that the p_{RR} increases as the sample size increases, and it decreases as variability increases.

Notice that when the coefficients of variation are 45%, and the total sample size is 40, the p_{RR} is only 48.5%. When a reference drug compared with itself can only have a 48.5% reproducibility probability, it is superfluous to follow the same study design to

Table 4: Biosimilarity index from R-R studies: p_{RR}

Sample size (n)	Coefficient of variation						
	20%	25%	30%	35%	40%	45%	50%
40	0.984	0.921	0.826	0.716	0.602	0.485	0.380
50	0.995	0.959	0.882	0.785	0.684	0.580	0.473
60	0.998	0.975	0.915	0.829	0.732	0.635	0.526
70	0.999	0.988	0.946	0.876	0.788	0.704	0.607
80	> 0.999	0.994	0.967	0.908	0.828	0.748	0.658
90	> 0.999	0.997	0.979	0.934	0.865	0.791	0.708
100	> 0.999	0.998	0.988	0.954	0.898	0.828	0.754

Figure 3: Biosimilarity index from R-R studies: p_{RR}



demonstrate that a ‘follow-on biologic’ is highly similar to this reference. There are two ways to correct a low p_{RR} : one is to choose a different criterion as in bioequivalence studies with the use of a scaled criterion; the other is to increase sample size. This also demonstrates the importance of conducting an R-R study: one is able to choose an appropriate criterion, sample size, or both, to establish reference standards.

An example

As shown above in the simulation studies, as coefficients of variation increase, it is less likely that similarity can be

declared even when R is compared with R. To further illustrate the effect of the high variability on the conclusion of biosimilarity, and how the biosimilarity index addresses the degree of similarity, we consider example data with a total sample size of 40 and coefficients of variation of 45%, again assuming no true mean difference.

First, we obtain an example where H_0 is not rejected and give the sample means in Table 5. The observed mean difference $\bar{Y}_{R_2} - \bar{Y}_{R_1} = 0.131$ and the observed variability $\hat{\sigma}_d = 0.368$. The 90% confidence interval of the mean difference is (-0.065, 0.327); similarity is not declared.

Under the same setting, another set of data is generated. Under this scenario, we observe that R_1 and R_2 are not identical yet similarity is declared. The sample means are given in Table 6, from which we could calculate the sample mean difference $\bar{Y}_{R_2} - \bar{Y}_{R_1} = -0.035$ and the variability $\hat{\sigma}_d = 0.279$. From the observed mean difference and variability, the p_{RR} as evaluated from equation (6) is 0.563. This example shows that when coefficients of variation is high, even though biosimilarity is declared based on the average similarity criterion, the degree of similarity as measured by biosimilarity index is low, so is the confidence that we have in the biosimilarity conclusion.

Conclusion

The simulation studies have shown that, as variance increases, the probability of declaring biosimilarity in the Step 1 test decreases; and, for those trials that pass the Step 1 test, the degree of similarity, quantified by biosimilarity index, decreases. The biosimilarity index is calculated for the trials that have passed

Table 5: Sample means from a study in which biosimilarity is not declared

Sequence	Period 1	Period 2	Sequence mean
1	$1R_1 = \bar{Y}_{.11} = 0.957$	$1R_2 = \bar{Y}_{.21} = 1.000$	$\bar{Y}_{.1} = 0.978$
2	$2R_2 = \bar{Y}_{.12} = 1.162$	$2R_1 = \bar{Y}_{.22} = 0.942$	$\bar{Y}_{.2} = 1.052$
Period mean	$\bar{Y}_{.1} = 1.060$	$\bar{Y}_{.2} = 0.970$	$\bar{Y}_{...} = 1.015$

Table 6: Sample means from a study in which biosimilarity is declared

Sequence	Period 1	Period 2	Sequence mean
1	$1R_1 = \bar{Y}_{.11} = 1.095$	$1R_2 = \bar{Y}_{.21} = 1.018$	$\bar{Y}_{.1} = 1.056$
2	$2R_2 = \bar{Y}_{.12} = 1.066$	$2R_1 = \bar{Y}_{.22} = 1.058$	$\bar{Y}_{.2} = 1.062$
Period mean	$\bar{Y}_{.1} = 1.081$	$\bar{Y}_{.2} = 1.038$	$\bar{Y}_{...} = 1.059$

the Step 1 test, thus it reflects the characteristics of the biological products that have already been declared biosimilar based on the average biosimilarity criterion. To apply the biosimilar index approach, an appropriate p_0 , see Table 1 for details of Step 3 in the testing procedure, needs to be determined. We recommend $p_0 = 0.7p_{RR}$ or $0.8p_{RR}$, such that the comparison is relative to the reference products. If we define $d = p_0/p_{RR}$, then d can be used to address the degree of similarity and the question of ‘how similar is highly similar’. Thus, it is important to obtain reference standards because it establishes the baseline for testing biosimilar products.

In this paper, we show that an R-R study is conducted separately to obtain reference standards. Reference standards can also be established concurrently using the three-arm study as proposed by Kang and Chow [2]. In the three-arm study, participants are randomly assigned to be treated with either test product (T) in one arm, or two different reference products (R_1, R_2) in the other two arms, such that reference standards are established concurrently in the two reference arms and used as baseline for the test product.

The idea as outlined in this paper can be applied to establish reference standards for biosimilarity index developed based on other criteria for the assessment of biosimilarity. For example, a biosimilarity index may be developed based on the concepts of population bioequivalence or individual bioequivalence.

Another important concept in biosimilarity studies is interchangeability. To assess interchangeability, it needs to be shown that the risk, i.e. safety or diminished efficacy of alternating or switching between use of the biosimilar and the reference product, are not greater than the risk of using the reference product without such alternation or switch. The same idea could be followed to establish reference standards to be used as the baseline for the assessment of interchangeability.

Acknowledgement

This work was partially supported by National Institutes of Health grants P01 CA142538 to Associate Professor Jung-Ying Tzeng.

Competing interests: None.

Provenance and peer review: Not commissioned; externally peer reviewed.

Co-authors

Associate Professor Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27607, USA

Professor Shein-Chung Chow, Professor of Biostatistics and Bioinformatics, Department of Bioinformatics and Biostatistics, Duke University School of Medicine, Suite 1102, Hock Plaza, 2424 Erwin Road, Durham, NC 27705, USA

References

1. U.S. Food and Drug Administration. Draft guidance for industry. Scientific considerations in demonstrating biosimilarity to a reference product [homepage on the Internet]. 2012 Feb [cited 2013 Jun 10]. Available from: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf>
2. Kang SH, Chow SC. Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Stat Med.* 2013;32(3):382-92.
3. Chow SC, Endreny L, Lachenbruch PA, Yang LY, Chi E. Scientific factors for assessing biosimilarity and drug interchangeability of follow-on biologics. *Biosimilars.* 2011;1:13-26.
4. Chow SC, Ju C. Assessing biosimilarity and interchangeability of biosimilar products under the Biologics Price Competition and Innovation Act. *Generics and Biosimilars Initiative Journal (GaBi Journal).* 2013;2(1):20-5. doi:10.5639/gabij.2013.0201.004
5. Hsieh TV, Chow SC, Yang LY, Chi E. The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics. *Stat Med.* 2013;32(3):406-13.
6. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinetic Biopharm.* 1987;15(6):657-80.
7. Shao J, Chow SC. Reproducibility probability in clinical trials. *Stat Med.* 2002; 21(12):1727-42.
8. Chow SC, Yang LY, Starr A, Chiu ST. Statistical methods for assessing interchangeability of biosimilars. *Stat Med.* 2013;32(3):442-8.
9. Chow SC, Liu PB. Design and analysis of bioavailability and bioequivalence studies. 3rd ed. New York: Taylor and Francis; 2008.

DOI: 10.5639/gabij.2013.0204.050

Copyright © 2013 Pro Pharma Communications International