HHS Public Access

Author manuscript

J Biopharm Stat. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

J Biopharm Stat. 2018 ; 28(2): 320–332. doi:10.1080/10543406.2017.1397012.

Improving the power to establish clinical similarity in a Phase 3 efficacy trial by incorporating prior evidence of analytical and pharmacokinetic similarity

Donglin Zeng[a], **Jean Pan**[b], **Kuolung Hu**[b], **Eric Chi**[b], and **D. Y. Lin**[a]

[a]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

[b]Amgen Inc, Thousand Oaks, CA, USA

Abstract

To improve patients' access to safe and effective biological medicines, abbreviated licensure pathways for biosimilar and interchangeable biological products have been established in the US, Europe, and other countries around the world. The US Food and Drug Administration and European Medicines Agency have published various guidance documents on the development and approval of biosimilars, which recommend a "totality-of-the-evidence" approach with a stepwise process to demonstrate biosimilarity. The approach relies on comprehensive comparability studies ranging from analytical and nonclinical studies to clinical pharmacokinetic/pharmacodynamic (PK/PD) and efficacy studies. A clinical efficacy study may be necessary to address residual uncertainty about the biosimilarity of the proposed product to the reference product and support a demonstration that there are no clinically meaningful differences. In this article, we propose a statistical strategy that takes into account the similarity evidence from analytical assessments and PK studies in the design and analysis of the clinical efficacy study in order to address residual uncertainty and enhance statistical power and precision. We assume that if the proposed biosimilar product and the reference product are shown to be highly similar with respect to the analytical and PK parameters, then they should also be similar with respect to the efficacy parameters. We show that the proposed methods provide correct control of the type I error and improve the power and precision of the efficacy study upon the standard analysis that disregards the prior evidence. We confirm and illustrate the theoretical results through simulation studies based on the biosimilars development experience of many different products.

Keywords

Bioequivalence; biological medicine; biosimilars; equivalence margins; rejection region; stepwise approach; totality of evidence

CONTACT Danyu Lin, lin@bios.unc.edu, Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lbps.

1. Introduction

A biological medicine is a large molecule derived from living cells (Dranitsaris et al., 2013). As the patents for a large number of biological products have already expired or are due to expire, there is an increasing interest from both the biopharmaceutical industry and the regulatory agencies in the development and approval of biosimilars (Noaiseh and Moreland, 2013; Ventola, 2013). The US Food and Drug Administration (FDA) defines a biosimilar as a product that is highly similar to the reference innovator product notwithstanding minor differences in clinically inactive components and with no clinically meaningful differences between the proposed biosimilar and the reference innovator product (US Food and Drug Administration, 2015a). Differences in cell lines and manufacturing processes make it impossible for biological products to be replicated exactly. Thus, the requirements to demonstrate similarity of biological products are different from those of demonstrating bioequivalence for generic small-molecule drug products that have an identical chemical structure.

To improve patients' access to safe and effective biological medicines, an abbreviated licensure pathway for biosimilar and interchangeable biological products was established under section 351 (k) of the Public Health Service Act in the US. Similar legal pathways for approval of biosimilars were established in the European and other countries around the world. Various guidance documents have been published by health authorities regarding the development and approval of biosimilars. The FDA (2015a, 2015b, 2015c, 2016a) and European Medicines Agency (2014) issued guidance documents that recommend a "totality-of-evidence" approach to assess biosimilarity with a stepwise process. The approach relies on comprehensive comparability studies with the reference product, progressing from analytical and nonclinical studies to clinical pharmacokinetic (PK) and pharmacodynamic (PD) studies (if there are relevant PD measures), and then to clinical efficacy studies. The analytical studies compare structural and functional characterization between the proposed biosimilar and the reference product. This serves as the foundation for a demonstration of biosimilarity. If there is residual uncertainty about biosimilarity after conducting analytical studies, animal testing, and clinical PK/PD studies, then clinical efficacy studies may be needed to adequately address that uncertainty, so as to support a demonstration that there are no clinically meaningful differences.

In light of the regulatory guidelines, it is strongly desirable to incorporate the prior knowledge of similarity or degree of uncertainty directly into the design and analysis of the clinical efficacy study. In general, the more similar the analytical and particularly the functional evaluations of the proposed biosimilar to the reference product are, the less residual uncertainty about biosimilarity there is to be addressed through the clinical studies. In addition, it is useful to incorporate PK similarity evidence to further reduce residual uncertainty and support a demonstration of no clinically meaningful differences.

In this article, we propose a statistical strategy that uses the analytical and PK similarity evidence from structural and functional characterization and phase 1 PK studies to reduce the sample size and enhance the power for a Phase 3 biosimilar efficacy study. Our assumption is that if the proposed biosimilar product and the reference product are highly

similar with respect to the analytical and phase 1 PK parameters, then they should also be similar with respect to the Phase 3 efficacy parameter. This strategy is in line with the FDA's guidance on the stepwise process and the totality-of-the-evidence approach to establish similarity. No such methods have been published, although there is a large body of literature on bioequivalence (e.g., Chow and Liu, 2008).

The rest of this article is organized as follows. In Section 2, we describe the proposed methods to incorporate the prior evidence of similarity (e.g., analytical and PK similarity) to improve the power of equivalence test and the precision of parameter estimation for a Phase 3 efficacy similarity study. We consider both the scenarios of a single source and multiple sources of prior similarity evidence. In Section 3, we conduct simulation studies to evaluate the proposed methods in realistic settings. Specifically, we show that the proposed methods preserve the type I error and enhance the power and estimation precision as opposed to disregarding the prior evidence when designing and analyzing the biosimilar clinical efficacy study. We also investigate the impact of the strength of analytical similarity on the proposed methods and show that combining the evidence from analytical and PK similarity can boost the power in the Phase 3 study. We provide some concluding remarks in Section 4.We relegate all theoretical details to appendices.

2. Methods

2.1. Incorporating single source of prior similarity evidence

Suppose that the parameter of interest in the prior evidence is θ, which may be a functional assessment for the mechanism of action, such as the primary receptor binding (particularly, tumor necrosis factor or TNF α binding) and antibody-dependent cellular cytotoxicity, in analytical studies or the logarithm of area under the curve (AUC) in the PK study. Suppose also that the parameter of interest in the Phase 3 trial is p, which is typically the response rate. Let (θ_T, p_T) and (θ_R, p_R) denote the values of (θ, p) for the biosimilar product to be tested and the reference product, respectively.

In the prior study, we test the null hypothesis

$$H_0^{(1)}: \theta_T - \theta_R < L' \text{ or } \theta_T - \theta_R > U'$$

against the alternative hypothesis

$$H_a^{(1)}: L' \le \theta_T - \theta_R \le U',$$

where $L' \le 0 \le U'$ are two specific margins. If, for example, θ is the logarithm of AUC, then $L' = -U' = \log(0.8)$. In the Phase 3 efficacy trial, we test the null hypothesis

$$H_0^{(3)}: \log \frac{p_T}{p_R} < L \text{ or } \log \frac{p_T}{p_R} > U$$

against the alternative hypothesis

$$H_a^{(3)}: L \leq \log \frac{p_T}{p_R} \leq U,$$

where $L \leq 0 \leq U$ are two specific margins. Typically, U and L are symmetric around 0, i.e., $U = -L$.

If there is strong empirical evidence that θ_T and θ_R are similar, then we wish to leverage this information when demonstrating the similarity between p_T and p_R in the Phase 3 trial. There are several challenges in formalizing this strategy. First, the parameters θ and p have different scales of measurement. Secondly, it is unclear how to efficiently incorporate the evidence about similarity from the prior study into the design and analysis of the Phase 3 study without making strong assumptions about the relationship between the parameters of the two studies. Lastly, because the prior evidence is empirical, we need to account for its randomness when incorporating it into the design and analysis of the Phase 3 study in order to control the overall type I error.

To address the first challenge, we rescale the parameters in the prior study and the Phase 3 study by defining a relative similarity measurement (RSM), which is the ratio between the absolute difference of the proposed biosimilar product and the reference product and the range of the margins. That is, the RSMs for the prior study and the Phase 3 study are defined as

$$RSM_1 = \frac{|\theta_T - \theta_R|}{|U' - L'|}, \; RSM_3 = \frac{|\log p_T - \log p_R|}{|U - L|},$$

respectively. Because RSM is a relative difference and thus has no unit, RSM_1 and RSM_3 are comparable. Furthermore, if the two products are similar within the margin, then RSM takes a value in [0, 1].

To address the second challenge, we impose a structural assumption on the relationship between RSM_1 and RSM_3, under which there exists a known positive constant c_1 such that

$$\text{if } RSM_1 < c_1, \text{ then } RSM_3 < \frac{\max\{|L|, |U|\}}{|U - L|}. \quad (1)$$

Thus, if the relative similarity in the prior study is within the bound c_1, then the difference between the two products in the Phase 3 study should be within the margins. In other words,

very similar performance between the two products in the parameter of the prior study renders evidence that the two products perform similarly with respect to the Phase 3 parameter. However, we do not specify any functional relationship between RSM_1 and RSM_3 but rather how the alternative hypothesis in the Phase 3 study is related to a bound for the relative similarity in the prior study. This structural assumption is minimal for the purpose of hypothesis testing. Clearly, the constant c_1 governs how much information is borrowed from the prior study. Specifying a reasonable c_1 requires some biological knowledge about the relationship between RSM_1 and RSM_3. For example, if we believe that RSM_1 is proportional to RSM_3, then for the case that $L = -U$, the structural assumption holds with $c_1 = RSM_1/(2RSM_3)$.

To address the last challenge, we must take into account the fact that there is some positive probability that the similarity evidence of $RSM_1 < c_1$ can be wrong when we construct the rejection region for hypothesis testing in the Phase 3 study. We propose to appropriately allocate the type I error as follows. Suppose that the overall type I error is set to α and that the null hypothesis that the two products are not similar holds for the Phase 3 study. We can reject the null hypothesis under one of two scenarios: (1) the empirical evidence in the prior study concludes that $RSM_1 < c_1$, such that the structural assumption leads to the conclusion that the two products are similar with respect to the Phase 3 parameter; and (2) the empirical evidence in the Phase 3 study indicates that the two products are similar. Thus, to control the overall type I error at α, we will control the error for case (1) to be under α_1 ($0 \le \alpha_1 \le \alpha$) while controlling the error for case (2) to be under ($\alpha - \alpha_1$). We provide a formal derivation and justification in Appendix A.

Note that, in the design stage, α_1 in the type I error spending needs to be specified before the Phase 3 study. We propose to search for the optimal α_1 such that the power under the alternative hypothesis in the Phase 3 study is maximized at the design stage. We show in Appendix B how to find this optimal α_1. After the Phase 3 study is completed, the ideas for using the structural assumption and error spending can be used to refine the confidence interval in the Phase 3 study at the analysis stage, where the error of using the prior evidence can be either fixed beforehand or determined in a data-adaptive manner. The details are provided in Appendix C.

2.2. Incorporating multiple sources of prior similarity evidence

We now extend the proposed methods to combine multiple sources of similarity evidence from, for example, analytical assessments and phase 1 PK studies. Suppose that there are K sources of similarity evidence. For $k = 1, \ldots, K$, the parameters for the proposed biosimilar product and the reference product in the kth source of evidence are denoted by θ_{Tk} and θ_{Rk}, respectively, and the corresponding margins are denoted by L'_k and U'_k. To combine the K sources of evidence, we define a weighted similarity metric

$$RSM_1 = \left| \sum_{k=1}^{K} w_k \frac{\theta_{Tk} - \theta_{Rk}}{U'_k - L'_k} \right|,$$

where $w_1, \ldots, w_K$ are pre-specified weights such that $\sum_{k=1}^{K} w_k = 1$. With this definition of RSM_1, the structural assumption is again given in (1).

If we define

$$\theta_T = \sum_{k=1}^{K} w_k \theta_{Tk}/(U'_k - L'_k),\ \theta_R = \sum_{k=1}^{K} w_k \theta_{Rk}/(U'_k - L'_k)$$

and set $L' = \leq 0.5$ and $U' = 0.5$, then $RSM_1 = |\theta_T - \theta_R|/|U' - L'|$. Thus, the design and analysis procedures described in section "Incorporating single source of prior similarity evidence" can be applied to multiple sources of prior evidence by treating θ_T and θ_R as the single parameter for the prior evidence. The implementation requires knowledge of the covariances between the estimators of $(\theta_{T1}, \ldots, \theta_{TK})^{\mathrm{T}}$ and $(\theta_{R1}, \ldots, \theta_{RK})^{\mathrm{T}}$ if the parameters are estimated from the same study. If the two sets of parameters pertain to analytical assessments and the PK study, then the covariance is zero.

The weights w_k reflect the importance of each source of prior evidence. If one of the w_k is 1 and the rest are 0, then we recover the set-up of one source of similarity evidence described in section "Incorporating single source of prior similarity evidence". The weights should be chosen on the basis of scientific knowledge, such as the extent of similarity in primary and secondary functions, together with overall structural/functional similarity. As functional tests are considered to be highly correlated with efficacy, more weights should be given to analytical similarity results than to PK similarity results. It is desirable to explore a range of values for the weights.

3 Numerical studies

3.1 Type I error and power

We conducted simulation studies to examine the performance of the proposed methods in practical situations. The first simulation study was designed to assess the type I error when the Phase 3 null hypothesis holds, and the second simulation study was designed to assess the power gain of the proposed method over the standard method that does not use the prior information. In both simulation studies, we simulated $n_1 = 50$ subjects from each treatment arm in the prior study and $n_3 = 300$ from each treatment arm in the Phase 3 study. We varied the threshold c_1 from 0.1 to 0.8 in a grid size 0.1. For each combination of simulation parameters, we set the number of replicates to 10,000.

In the first simulation study, we set the parameter value for the reference product in the prior study as $\theta_R = 5$ and set the two margins as $L' = -U' = \log(0.8)$. In addition, we set the response rates for the two products in Phase 3 as $p_R = 0.4$ and $p_T = 0.75p_R$. We let the two margins for Phase 3 be $L = -U = \log(0.75)$, such that the Phase 3 null hypothesis holds. Since the structural assumption must hold, the parameter value for the proposed biosimilar product, θ_T, should satisfy $RSM_1 \geq c_1$. We particularly chose $RSM_1 = 1.25c_1$ by setting $\theta_T = 1.25c_1(U' - L') + \theta_R$. Thus, we generated n_1 measurements from $N(\theta_R, 0.11^2)$ for the

reference product and from $N(\theta_T, 0.1^2)$ for the proposed biosimilar product. In Phase 3, we generated n_3 binary responses from $Bernoulli(p_R)$ for the reference product and from $Bernoulli(p_T)$ for the proposed biosimilar product.

In the second simulation study, we adopted the same simulation set-up but let $p_T = p_R = 0.4$ and θ_T be $0.75c_1(U' - L') + \theta_R$. The latter condition guarantees that the structural assumption is satisfied under the alternative hypothesis $H_a^{(3)}$. With $n_3 = 300$, the power of the Phase 3 trial is 80% at $\alpha = 0.05$.

For each simulated data set, we applied the method described in Appendix A to obtain the rejection region at the significance level of $\alpha = 5\%$. We set α_1/α to 0, 0.2, or 0.4. Clearly, $\alpha_1 = 0$ is equivalent to using the Phase 3 data only. We also applied the method described in Appendix B with 1 million Monte Carlo samples to obtain the optimal type I error for using the prior evidence, α_1^{opt}. Finally, we used the method described in Appendix C to construct the 90% confidence intervals for $\log(p_T/p_R)$ corresponding to different choices of α_1.

Figure 1 displays the results from the first simulation study, including the empirical type I error at the 5% significance level, the coverage probability of the 90% confidence interval, and the relative width of the 90% confidence interval based on different α_1 over that of using the Phase 3 data only. Clearly, the type I error rates for all the methods are below the nominal significance level for all values of c_1, and the coverage probabilities of the confidence intervals are all above the nominal level of 90%. Using a pre-specified value of α_1 tends to yield conservative type I error and wider confidence intervals. By contrast, the proposed method based on the optimal choice of α_1 provides accurate control of the type I error and the tightest confidence intervals. For the particular parameter set-up in this simulation, the method based on α^{opt} performs very similarly to the method using the Phase 3 data only because the value of α^{opt} turns out to be close to 0. This is due to the fact that $p_T = p_R$, such that the Phase 3 data almost possess sufficient power to reject the null hypothesis.

Figure 2 displays the results from the second simulation study. The proposed method based on α^{opt} has the highest power and the shortest confidence intervals for all choices of c_1. The power increases and the confidence interval becomes narrower as c_1 increases. Thus, using the prior evidence through the structural assumption in an optimal way leads to higher power and more accurate estimation for the Phase 3 study.

3.2 Impact of the strength of analytical similarity evidence

It is worthwhile to investigate how the strength of the prior evidence may impact the Phase 3 study in our approach. The strength mostly depends on the extent of similarity in the prior study. Thus, we conducted a simulation study to evaluate the power and estimation results for the Phase 3 study after incorporating the evidence from analytical assessments, where the extent of similarity varies in terms of the variability or the difference pertaining to a functional test (e.g., TNFα binding affinity) between the proposed biosimilar product and the reference product. Specifically, we generated 10 measurements from $N(1.08 + \delta, \sigma^2)$ for the reference product and 10 measurements from $N(1.08, \sigma^2)$ for the proposed biosimilar product. Clearly, the larger the value of δ or σ is, the weaker the evidence is. We let $L' = -U$

′ = −0.1497. For Phase 3, we set $p_T = p_R = 0.85$ and set the margins for $p_T - p_R$ as $L = -U = -0.15$. We set the type I error at 0.05 and restricted α_1 to be less than 0.01 so as to avoid the situation in which the analytical evidence is sufficient to conclude the success of Phase 3. We first fixed $\delta = 0.04$ and varied σ in order to assess the impact of the variability of the measurements. We then fixed $\sigma = 0.09$ and varied δ in order to assess the impact of the actual product difference in analytical assessments.

If the prior evidence is not used, then the sample size of 98 per arm is required to achieve the power of 80%. Figure 3(a) displays the ratio of the sample size for using the analytical evidence to the sample size for not using the prior evidence (i.e., 98) under five different values of σ. To determine the sample size, we searched over a wide range of sample sizes to obtain the smallest sample size such that the power calculated in Appendix B was at least 80%. Figure 3(b) displays the power of the proposed method for analyzing the Phase 3 data. The strength of the analytical evidence, characterized by σ, has strong impact on the results. For example, under $c_1 = 0.4$, the value of $\sigma = 0.07$ reduces the sample size from 98 to about 69, and even with the reduced sample size, the average power is increased from 80% to 87%. Figure 3(c) and (d) show the ratio of the sample sizes and the power when the difference between the two products, δ, changes. The conclusion is similar: the smaller the value of δ is (i.e., the stronger the analytical similarity is), the more gain we achieve for both the design and the analysis of the Phase 3 study.

3.3 Impact of analytical and PK similarity evidence

In this simulation study, we combined two sources of similarity evidence, one from analytical assessments and one from a phase 1 PK study, and evaluated their impact on the sample size and power of a future Phase 3 efficacy study. For the first source, we generated comparative measurements from $N(1.08, 0.089^2)$ for the proposed biosimilar product and from $N(1.12, 0.1^2)$ for the reference product based on the established functional similarity (US Food and Drug Administration, 2016b). The corresponding margins were $L_1' = -U_1' = -0.1497$. The parameter in the second source is the logarithm of the AUC in a PK study (Yoo et al., 2017). We generated 45 observation on the logarithm of the AUC for the reference product from $N(8.9099, 0.3^2)$ and 96 observations for the proposed biosimilar product from $N(8.9668, 0.3^2)$, and we set $L_2' = -U_2' = \log(0.8)$. For the Phase 3 study, we set $p_T = p_R = 0.85$, and the margins of their equivalence were $L = -U = -0.15$. We used the proposed method to combine these two sources of similarity evidence. The two sources were derived from independent studies, so the parameter estimators are uncorrelated. We varied the weight for the analytic evidence, i.e., w_1, from 1 to 0, in order to examine the impact of the contribution from each evidence. Again, we set α_1 to be at most 0.01. As shown in Figure 4, adding PK similarity evidence (i.e., setting $w_1 < 1$) further reduces the sample size and increases the study power as compared to using analytical similarity evidence alone (i.e., $w_1 = 1$). When $c_1 = 0.4$ and the desired power is 0.85, the candidate values of w_1 are 0.25, 0.5, and 0.75. Since the analytic similarity is deemed the most relevant to the clinical outcome, the best choice of the weight would be $w_1 = 0.75$.

4 Discussion

We have developed a simple and effective strategy to reduce the sample size and improve the power and estimation for the Phase 3 efficacy trial of biosimilars by leveraging the prior similarity evidence from multiple sources such as analytical assessments and PK studies. Our approach hinges on the structural assumption, under which strong evidence for similarity in the parameters of prior studies implies similarity in the Phase 3 parameter. This assumption is a qualitative rather than a functional relationship between the two similarity measures.

From a Bayesian point of view, the structural assumption can be regarded as a prior distribution for the two similarity measures. The corresponding prior is qualitative. We may use a more continuous prior distribution by assuming that $RSM_3/RSM_1 - 1 \sim N(0, \sigma^2)$, which is equivalent to a ridge penalty for $(RSM_3/RSM_1 - 1)^2$. The hyperparameter σ^2 plays the role of c_1 to govern how much prior similarity evidence is used to reinforce the Phase 3 trial. The continuous prior/penalty is computationally easier to handle than the discrete prior. However, it is not desirable to constrain the parameters when RSM_3 is far from RSM_1.

The choice of c_1 is a critical aspect of the proposed methods. We recommend to determine c_1 using the evidence on related products with approved indications. For example, for an infliximab biosimilar product approved by the FDA in 2016, the analytical similarity (as measured by TNFα binding affinity) yields the difference between the reference and test products as 2.59% with the margins of −7.04% and 7.04%, and the clinical similarity as measured by ACR20 yields the difference of 2% with the margins of −15% and 15% (Yoo et al., 2017). For a biosimilar later approved by the FDA for similar indications, the analytical similarity is −3.6% with the margins of −14.97% and 14.97%, and the clinical similarity is −0.4% with the margins of −12% and 12% (US Food and Drug Administration, 2016b). Thus, the ratio RSM_1/RSM_3 is estimated at 2.76 for the first study and at 7.46 for the second study. This empirical evidence suggests that c_1, which is chosen as half of this ratio to satisfy the structural assumption, can be as large as 1.3. In addition to empirical evidence, it is worthwhile to conduct sensitivity analysis to examine the power over a range of values for c_1 when designing a new Phase 3 study and when analyzing data after the study is completed.

If the prespecified value of c_1 is not small, the prior evidence may be strong enough to reject the Phase 3 null hypothesis, such that there is no need to run the Phase 3 study. To reduce the likelihood of this scenario, we may, as in the numerical studies, restrict a_1 to be within a certain threshold, such that we will use only the prior evidence when $RSM_1 < c_1$ is supported by prior studies with very high probability. The optimal error spending may be different between the design stage and the analysis stage, as the former aims to maximize power whereas the latter aims to construct the narrowest confidence interval that depends on the empirical data in the Phase 3 study. This difference was observed in our numerical studies.

We have assumed that the endpoint for the Phase 3 efficacy study is a binary response. We can easily extend our methods to a time-to-event endpoint. The null hypothesis will then pertain to the hazard ratio instead of the ratio of the two response rates.

The FDA recommended a tier approach to statistical analysis based on a critically risk ranking as described by Tsong et al. (2017). For the attributes with the highest risk to clinical outcomes (Tier 1), which include assays that evaluate clinically relevant mechanisms of action of the product, a demonstration of statistical equivalence is required. Specifically, the equivalence is demonstrated if the confidence interval for the difference in the mean between the proposed biosimilar product and the reference product is fully contained within the equivalence acceptance region. For the quality attributes with lower risk ranking (Tier 2), the similarity is assessed by comparing the individual results of the proposed biosimilar product with a quality range based on the mean and standard deviation of the reference product dataset. Finally, for the quality attributes with the lowest risk ranking and those that do not deliver quantitative results (Tier 3), the similarity is assessed qualitatively by using raw data and graphical comparisons.

The methods proposed in this article can be applied directly to the quantitative similarity evidence from Tier 1 quality attributes, as illustrated in sections “Impact of the strength of analytical similarity evidence” and “Impact of analytical and PK similarity evidence”. The proposed methods can also be applied to the similarity evidence from Tier 2 attributes by treating the quality range as the margin. However, the requirement to show similarity for Tier 2 quality attributes is different from the rigorous bio-equivalence test described in this article, so the use of the proposed methods to incorporate the similarity evidence from Tier 2 attributes may require further consideration. For Tier 3 quality attributes whose similarity evidence is qualitative, the proposed methods cannot be directly applied. Since Tier 2 and Tier 3 quality attributes have low or no risk to patients, the proposed methods remain a viable approach to efficiently leverage the prior analytical similarity information from Tier 1 attributes for the design and analysis of a biosimilar clinical study.

In the proposed methods, weights are introduced to combine multiple sources of similarity evidence. As with any weighted statistical methods, the choice of the weights is challenging, and the weights should reflect the importance of each source of the prior evidence. Since functional attributes are most relevant to clinical outcomes, we recommend that more weights be given to the functional similarity evidence, especially the Tier 1 functional attributes. In addition, the weights should be reference-drug specific and be determined according to the scientific knowledge about the relevance of the prior evidence (from structural, functional, and PK studies) on the clinical outcomes for the reference product. Ultimately, the acceptable values for weights become a negotiating point between the sponsor and regulatory agencies, and it will be up to regulatory agencies to decide whether the same approach to design a clinical similarity study should be implemented by different biosimilar applicants.

Acknowledgments

This article is dedicated to the memory of Eric Chi. The authors would like to thank Dr. Richard Markus for his helpful discussions and comments. They would also like to thank two reviewers for their constructive comments.

References

Chow, S-C., Liu, JP. Design and Analysis of Bioavailability and Bioequivalence Studies. 3. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series; 2008.

Dranitsaris G, Dorward K, Hatzimichael E, Amir E. Clinical trial design in biosimilar drug development. Investigational New Drugs. 2013; 31:479–487. [PubMed: 23161336]

European Medicines Agency. Guideline on Similar Biological Medicinal Products. London, UK: CHMP; 2014. Report No. CHMP/437/04 rev 1

Noaiseh G, Moreland L. Current and future biosimilars: Potential practical applications in rheumatology. Dove Medical Press. 2013; 3:27–33.

Tsong Y, Dong X, Shen M. Development of statistical methods for analytical similarity assessment. Journal of Biopharmaceutical Statistics. 2017; 27(2):192–205.

US Food and Drug Administration. Scientific considerations in demonstrating biosimilarity to a reference product: Guidance for industry. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); Rockville, MD: 2015a.

US Food and Drug Administration. Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product: Guidance for industry. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); Rockville, MD: 2015b.

US Food and Drug Administration. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). Silver Spring, MD: U.S. Food and Drug Administration; 2015c. Biosimilars: Questions and answers regarding implementation of the biologics price competition and innovation act of 2009.

US Food and Drug Administration. Clinical pharmacology data to support a demonstration of biosimilarity to a reference product. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); Rockville, MD: 2016a.

US Food and Drug Administration. Arthritis Advisory Committee Meeting Briefing Book for BLA 761024. Silver Spring, MD: U.S. Food and Drug Administration; 2016b.

Ventola CL. Biosimilars: Part 1: Proposed regulatory criteria for FDA approval. Pharmacy and Therapeutics. 2013; 38:270–287. [PubMed: 23946620]

Yoo DH, Suh CH, Shim SC, et al. A multicentre randomised controlled trial to compare the pharmacokinetics, efficacy and safety of CT-P10 and innovator rituximab in patients with rheumatoid arthritis. Ann Rheum Dis. 2017; 76:566–570. [PubMed: 27624791]

Appendix A. Rejection Region for Hypothesis Testing in the Phase 3 Study

First, we construct a rejection region based on the prior evidence data such that the type I error to reject the null hypothesis $H_0^*: RSM_1 \geq c_1$ is controlled at α_1. Specifically, let $\hat{\theta}_T$ and $\hat{\theta}_R$ denote the estimators of θ_T and θ_R, and let $\hat{\sigma}_T^2$ and $\hat{\sigma}_R^2$ denote the corresponding variance estimators. Write $\delta_1 = \theta_T - \theta_R$ and $\hat{\delta}_1 = \hat{\theta}_T - \hat{\theta}_R$. Then $\hat{\delta}_1$ is approximately normal with mean δ_1 and variance $\hat{v}_1^2 = \hat{\sigma}_T^2 + \hat{\sigma}_R^2$. Thus,

$$P\left(\hat{\delta}_1 - \hat{v}_1 z_{1-\alpha_1/2} \leq \delta_1 \leq \hat{\delta}_1 + \hat{v}_1 z_{1-\alpha_1/2}\right) = 1 - \alpha_1,$$

where z_p is the $100p$ th percentile of the standard normal distribution. It follows that

$$P(|\delta_1| \le \hat{Z}_{\alpha_1}) \ge 1-\alpha_1,$$

where

$$\hat{Z}_{\alpha_1} = \max\left(\left|\hat{\delta}_1 - \hat{v}_1 z_{1-\alpha_1/2}\right|, \left|\hat{\delta}_1 + \hat{v}_1 z_{1-\alpha_1/2}\right|\right). \quad (2)$$

That is, $[0, \hat{Z}_{\alpha_1}/|U' - L'|]$ is a $(1-\alpha_1)100\%$ confidence interval for RSM_1. Hence, we reject H_0^* when c_1 is within this interval, i.e., $\hat{Z}_{\alpha_1} \le c_1|U' \le L'|$. This rejection region has the type I error of α_1.

Second, we construct a rejection region based on the Phase 3 data such that the type I error is controlled at α_3. This can be achieved by using the standard equivalence test. Define $\delta_3 = \log p_T/p_R$ and $\hat{\delta}_3 = \log \hat{p}_T/\hat{p}_R$. We estimate the variance of $\hat{\delta}_3$ by $\hat{v}_3^2 = \hat{p}_T(1-\hat{p}_T)/(n_T\hat{p}_T^2) + \hat{p}_R(1-\hat{p}_R)/(n_R\hat{p}_R^2)$, where n_T and n_R are the sample sizes for the proposed biosimilar product and the reference product, respectively, in the Phase 3 study. Clearly, $\hat{\delta}_3$ is approximately normal with mean δ_3 and variance $\hat{v}_3^2$. Thus, a $(1-2\alpha_3)100\%$ confidence interval for δ_3 is$[\hat{L}_{\alpha_3}, \hat{U}_{\alpha_3}]$, where

$$\hat{L}_{\alpha_3} = \hat{\delta}_3 - \hat{v}_3 z_{1-\alpha_3}, \hat{U}_{\alpha_3} = \hat{\delta}_3 + \hat{v}_3 z_{1-\alpha_3}.$$

Hence, we reject $H_0^{(3)}$ if $[\hat{L}_{\alpha_3}, \hat{U}_{\alpha_3}]$ is contained in $[L,U]$. This rejection region controls the type I error at the level of α_3. Combining the above two regions and setting $\alpha_3 = \alpha - \alpha_1$, we obtain the overall rejection rule for $H_0^{(3)}$

$$A_{\alpha_1} = I([\hat{L}_{\alpha-\alpha_1}, \hat{U}_{\alpha-\alpha_1}] \subset [L,U] \text{ or } \hat{Z}_{\alpha_1} \le c_1 \,|\, U'-L'|).$$

Suppose that the null hypothesis $H_0^{(3)}$ holds. Then the structural assumption implies that the null hypothesis H_0^* also holds. Thus,

$$\begin{aligned} P(A_{\alpha_1}=1) &\le P([\hat{L}_{\alpha-\alpha_1}, \hat{U}_{\alpha-\alpha_1}] \subset [L,U]) + P(\hat{Z}_{\alpha_1} \le c_1 \,|\, U'-L'|) \\ &\le (\alpha-\alpha_1)+\alpha_1 = \alpha. \end{aligned}$$

In other words, the overall probability of rejection is no larger than α under $H_0^{(3)}$.

Appendix B. Optimal Type I Error Spending

Under the alternative hypothesis $H_a^{(3)}$, the probability of no rejection is

$$P(L > \hat{L}_{\alpha-\alpha_1} \text{ or } \hat{U}_{\alpha-\alpha_1} > U)P(\hat{Z}_{\alpha_1} \geq c_1 \mid U' - L' \mid)$$
$$\leq \left\{\Phi\left(\frac{L-\delta_3}{v_3} - z_{\alpha-\alpha_1}\right) + \Phi\left(-\frac{U-\delta_3}{v_3} - z_{\alpha-\alpha_1}\right)\right\}$$
$$\times P\left(\max\left\{\left|N(0,1) - z_{1-\alpha_1/2} + \frac{\delta_1}{v_1}\right|, \left|N(0,1) + z_{1-\alpha_1/2} + \frac{\delta_1}{v_1}\right|\right\} \geq \frac{c_1 \mid U' - L' \mid}{v_1}\right),$$

where $N(0, 1)$ denotes a standard normal random variable, Φ is the standard-normal distribution function, and v_1 and v_3 are obtained from $\hat{v}_1$ and $\hat{v}_3$, respectively, by replacing the parameter estimators with the true parameter values. Denote the right side of the above inequality by $G(a_1)$. We then search over $a_1 \in [0, a]$ such that the power $1 - G(a_1)$ is maximized. Note that $1 - G(0)$ is the power without using the prior evidence.

We construct the optimal rejection region as follows:

Step 1. We obtain $\hat{\theta}_T$, $\hat{\theta}_R$ and $(\hat{\delta}_1, \hat{v}_1)$ using the prior study data.

Step 2. We obtain $\delta_3 = \log p_T/p_R$ and v_3 under $H_a^{(3)}$.

Step 3. We search over a grid of a_1 in $[0, a]$ to evaluate $G(a_1)$, where δ_1 and v_1 are replaced by $\hat{\delta}_1$ and $\hat{v}_1$, respectively. In particular, the probabilities in $G(a_1)$ are calculated by Monte Carlo simulation.

Step 4. We determine α_1^{opt} which maximizes $1 - G(a_1)$, such that we reject $H_0^{(3)}$ using $A_{\alpha_1^{opt}}$.

Appendix C. Refined Confidence Region in the Phase 3 Study Using the Prior Evidence

The idea of using the prior information for hypothesis testing in the Phase 3 study can be extended to obtain a narrower confidence interval for δ_3. To construct a $(1 - 2a)100\%$ confidence interval (corresponding to the hypothesis testing for equivalence with the type I error a), we split a into a_1 and $(a - a_1)$. Using the prior evidence, we calculate $\hat{Z}_{a_1}$ as before such that $P(|\delta_1| \leq \hat{Z}_{a_1}) \geq 1 - a_1$. Thus, if $RSM_1 \geq c_1$, then

$$P(\hat{Z}_{\alpha_1} \leq c_1 \mid U' - L' \mid) \leq P(\hat{Z}_{\alpha_1} \leq \mid \delta_1 \mid) \leq \alpha_1.$$

Using the Phase 3 data, we construct a $(1 - 2a + 2a_1)100\%$ confidence interval for δ_3 as $(\hat{L}_{a-a_1}, \hat{U}_{a-a_1})$, such that

$$P(\delta_3 \le \hat{L}_{\alpha-\alpha_1}) \le \alpha-\alpha_1, (\delta_3 \ge \hat{U}_{\alpha-\alpha_1}) \le \alpha-\alpha_1 .$$

By combining the above two confidence intervals, we obtain

$$[\hat{L}_f, \hat{U}_f] = \begin{cases} [\hat{L}_{\alpha-\alpha_1}, \hat{U}_{\alpha-\alpha_1}] & if\ \hat{Z}_{\alpha_1} > c_1 \mid \hat{U} - L' \mid \\ [\hat{L}_{\alpha-\alpha_1} \vee L, \hat{U}_{\alpha-\alpha_1} \wedge U] & otherwise \end{cases}$$

where $a \vee b = \max(a, b)$, and $a \wedge b = \min(a, b)$. If $\alpha_1 = 0$, then $[\hat{L}_f, \hat{U}_f]$ reduces to $[\hat{L}_\alpha, \hat{U}_\alpha]$, which is the confidence interval based on the Phase 3 data only.

To see why the proposed confidence interval has the correct coverage, we note that

$$\begin{aligned} P(\delta_3 \le \hat{L}_f) &= P(\delta_3 \le \hat{L}_{\alpha-\alpha_1}, \hat{Z}_{\alpha_1} > c_1 \mid U' - L' \mid) + P(\delta_3 \le \hat{L}_{\alpha-\alpha_1} \vee L, \hat{Z}_{\alpha_1} \le c_1 \mid U' - L' \mid) \\ &\le (\alpha-\alpha_1) P(\hat{Z}_{\alpha_1} > c_1 \mid U' - L' \mid) + P(\delta_3 \le \hat{L}_{\alpha-\alpha_1} \text{ or } \delta_3 \le L) P(\hat{Z}_{\alpha_1} \le c_1 \mid U' - L' \mid) . \end{aligned}$$

We consider the second term on the right side. If $\delta_3 > L$, then this probability is less than

$$P(\delta_3 \le \hat{L}_{\alpha-\alpha_1}) P(\hat{Z}_{\alpha_1} \le c_1 \mid U' - L' \mid) \le (\alpha-\alpha_1) P(\hat{Z}_{\alpha_1} \le c_1 \mid U' - L' \mid) .$$

If $\delta_3 \le L$, then since the structural assumption implies that $RSM_1 > c_1$ and thus $|\delta_1| > \hat{Z}_{\alpha_1}$, this probability is less than

$$P(\hat{Z}_{\alpha_1} \le c_1 \mid U' - L' \mid) \le P(\hat{Z}_{\alpha_1} \le \mid \delta_1 \mid) \le \alpha_1 .$$

In either case,

$$P(\delta_3 \le \hat{L}_f) \le \alpha .$$

Likewise, we conclude that $P(\delta_3 \ge \hat{U}_f) \le \alpha$. That is, $[\hat{L}_f, \hat{U}_f]$ is a valid $(1 - 2\alpha)$-confidence interval for δ_3. We can search for the optimal α_1 such that the resulting confidence interval has the shortest length.

Remark 1

If we use the confidence interval to perform the hypothesis test by rejecting $H_0^{(3)}$ when $[\hat{L}_f, \hat{U}_f] \subset [L, U]$, then we reject $H_0^{(3)}$ when either $\hat{Z}_{a_1} \leq c_1|U' - L'|$ or $[\hat{L}_{a-a_1}, \hat{U}_{a-a_1}] \subset [L, U]$. This is exactly the rejection region described in Appendix A.

Remark 2

When analyzing the Phase 3 data, we may adjust for covariates through a log-linear regression model:

$$\log p = \delta_3 G + \xi^{\mathrm{T}} X,$$

where G is the indicator for the proposed biosimilar product versus the reference product, and X is the set of covariates including the unit component.

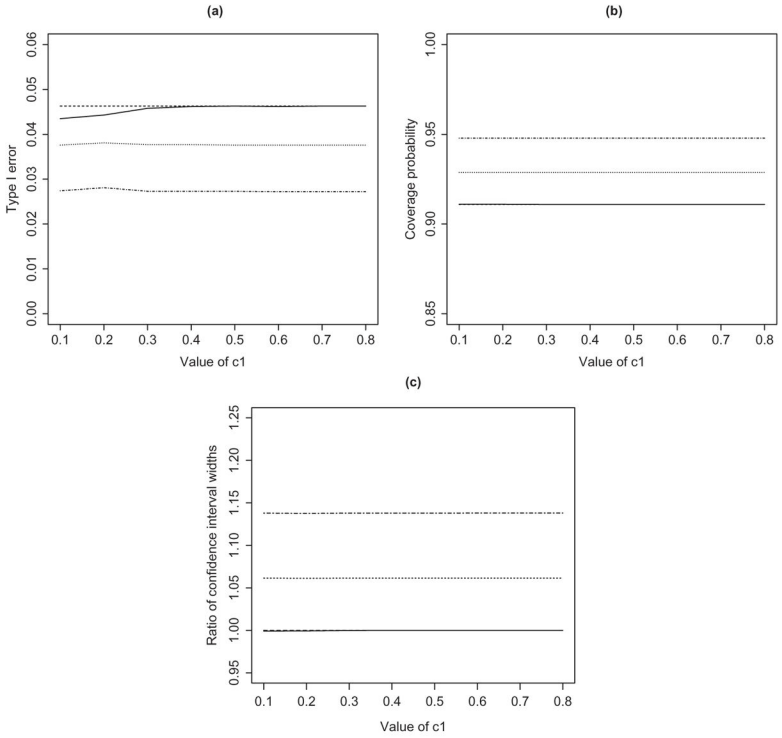

Figure 1.

Simulation results under the null hypothesis: (a) type I error of the equivalence test; (b) coverage probability of the 90% confidence interval; and (c) ratio of the widths of the confidence intervals with versus without the prior evidence. The solid, dashed, dotted, and dot-dashed curves pertain to α_1^{opt}, $a_1 = 0$, $a_1 = 0.2a$, and $a_1 = 0.4a$, respectively. The solid and dashed curves are indistinguishable.

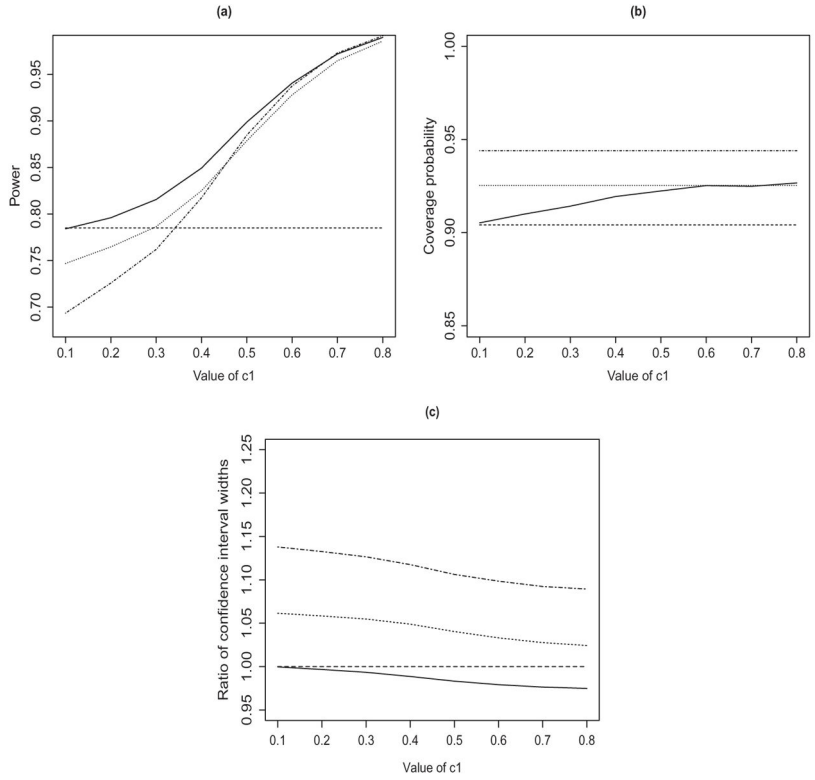

Figure 2.

Simulation results under the alternative hypothesis: (a) power of the equivalence test; (b) coverage probability of the 90% confidence interval; and (c) ratio of the widths of the confidence intervals with versus without the prior evidence. The solid, dashed, dotted, and dot-dashed curves pertain to α_1^{opt}, $a_1 = 0$, $a_1 = 0.2a$, and $a_1 = 0.4a$, respectively.

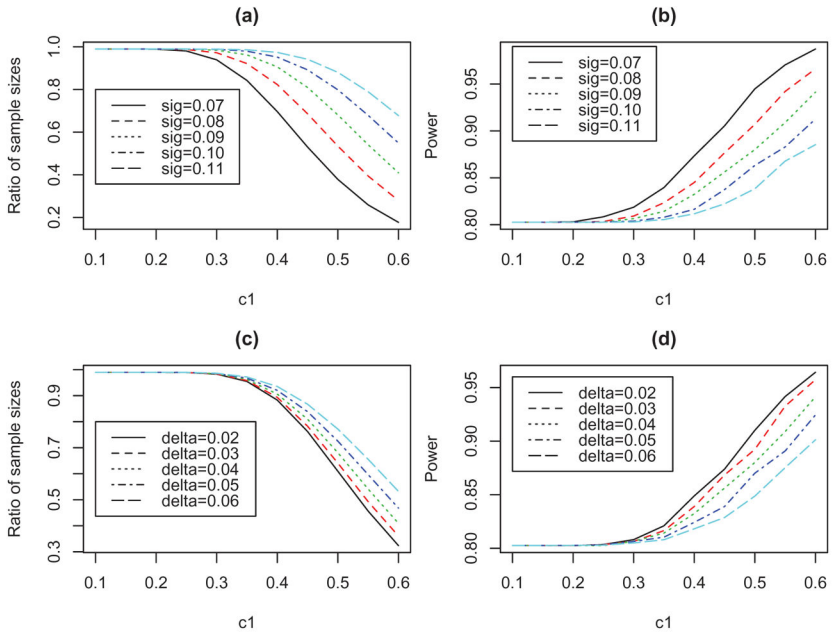

Figure 3.

Simulation results for using the analytical assessment evidence in both designing and analyzing the Phase 3 study: (a) and (c) show the ratio of the sample sizes in the Phase 3 study to achieve the power of 80% when the prior evidence is used versus when it is not used; and (b) and (d) show the power of the proposed equivalence test in analyzing the Phase 3 data.

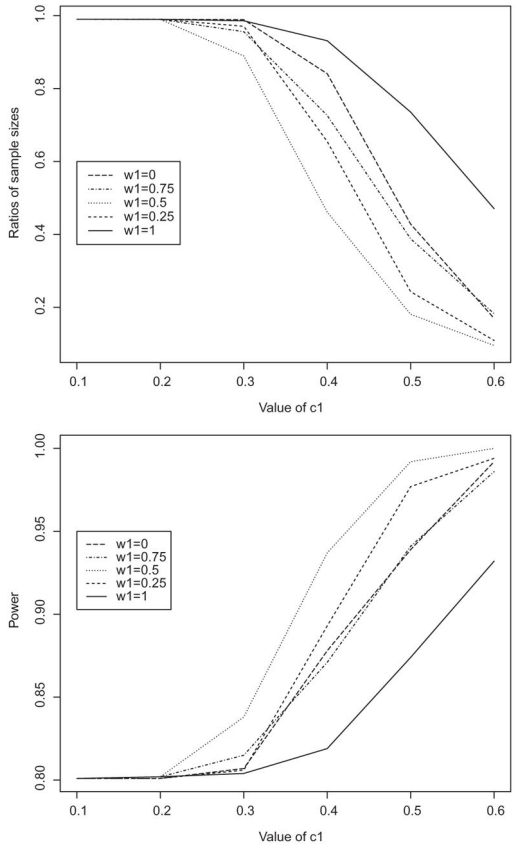

Figure 4.

Simulation results for combining both analytical and PK similarity evidence: (a) ratio of the sample sizes in the Phase 3 study to achieve the power of 80% when the prior evidence is used versus when it is not used; and (b) power of the proposed equivalence test in analyzing the Phase 3 data.