Statistics
in Medicine

Research Article

Received 7 November 2016, Accepted 8 April 2017 Published online 11 May 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7326

On assessing bioequivalence and interchangeability between generics based on indirect comparisons

Jiayin Zheng,[a,b*†] Shein-Chung Chow[a,b] and Mengdie Yuan[c]

As more and more generics become available in the market place, the safety/efficacy concerns may arise as the result of interchangeably use of approved generics. However, bioequivalence assessment for regulatory approval among generics of the innovative drug product is not required. In practice, approved generics are often used interchangeably without any mechanism of safety monitoring. In this article, based on indirect comparisons, we proposed several methods to assessing bioequivalence and interchangeability between generics. The applicability of the methods and the similarity assumptions were discussed, as well as the inappropriateness of directly adopting adjusted indirect comparison to the field of generics' comparison. Besides, some extensions were given to take into consideration the important topics in clinical trials for bioequivalence assessments, for example, multiple comparisons and simultaneously testing bioequivalence among three generics. Extensive simulation studies were conducted to investigate the performances of the proposed methods. The studies of malaria generics and HIV/AIDS generics prequalified by the WHO were used as real examples to demonstrate the use of the methods.

Keywords: average bioequivalence; drug interchangeability; generics; geometric mean ratio (GMR); indirect comparison

1. Introduction

Average bioequivalence assessment for regulatory review and approval between a generic and a brand name drug is well established and widely used in pharmaceutical industry [1–5]. Before a generic can be approved by the United States Food and Drug Administration (FDA), the sponsor is required to conduct a bioequivalence study to demonstrate that the generic is bioequivalent to the brand name drug in terms of the rate and extent of drug absorption. Under the fundamental bioequivalence assumption, two drug products are considered therapeutic equivalent if they are shown to be bioequivalent in drug absorption profile. An approved generic can then be used as a substitute of the brand name drug. In particular, the free interchangeability is generally assumed for small-molecule generics, resulting in the substitution of various generics in pharmacies. Because small-molecule drugs are generally safe and effective, these assumptions and practices usually work. However, important exceptions exist.

When several generics are bioequivalent to the same brand name drug, it is not obvious that they are bioequivalent to each other. Anderson and Hauck [6] demonstrated that although with two or three generics, the probability of such lack of bioequivalence is fairly low, the concern becomes substantial with five, six, or more products on the market. As more and more generics become available in the market place, and the fact that approved generics may be used interchangeably, it is a concern whether (1) the quality of approved generics is consistent and (2) the interchangeability among the approved generics is safe and efficacious. For example, with respect to the pharmacokinetic mean responses, if an approved generic

[a]*National Clinical Research Center for Digestive Disease, Beijing Friendship Hospital, Capital Medical University, Beijing, China*
[b]*Duke University School of Medicine, 2424 Erwin Road, Hock Plaza, Suite 1102, Durham, NC 27705, U.S.A.*
[c]*Center for Drug Evaluation and Research Organization, Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
[*]*Correspondence to: Jiayin Zheng, Duke University School of Medicine, 2424 Erwin Road, Hock Plaza, Suite 1102, Durham, NC 27705, U.S.A.*
[†]*E-mail: statzjy@gmail.com*

A falls on the bioequivalence lower end (say 80%) of the brand name drug and another approved generic B is on the higher end (say 120%), the relative change in response of a subject who switches from A to B could lead to a drastic change (say 50%) in blood concentration, which could potentially raise a safety concern. Also, strong safety concerns were repeatedly raised about the generic-to-generic substitution of, for example, antiepileptics [7, 8] and immunosuppressant products [9]. However, bioequivalence assessment among generics is not required by the regulatory agencies such as FDA [2, 10], and there is often a lack of head-to-head comparative trials between all available generics of the same brand name drug to ensure the interchangeability between them.

As the settings of the original bioequivalence trials are not identical, using the original data directly to test the bioequivalence between generics is not recommended. To deal with this difficulty, the following two approaches may be useful: (1) Calibrate the original data based on the summary statistics of the brand name drug from both bioequivalence trials, and then analyze the calibrated data; (2) use the results from different trials to estimate the relative bioavailabilities between generics through indirect comparison. The latter can be an alternative and valuable way. Following this idea, several approaches of indirect comparison in investigating the bioequivalence between generics have been employed and reviewed [11–14]. Among them, the adjusted indirect comparison was considered by Garcia-Arieta *et al.* [12] as the simplest and most suitable method for bioequivalence studies. The adjusted indirect comparison is a useful tool for assessing the bioequivalence between two generics of the same brand name drug and partly preserves the power of randomized controlled trials. But nonetheless, it requires the original studies be sufficiently powered ($> 80\%$) and the point estimate difference between generics not exceed the 7% difference [12]. To get a better performance, Garcia-Arieta *et al.* [12] also extended the acceptance limit from $\pm 20\%$ to $\pm 30\%$ range, which seems arbitrary.

For clinical practice, the bioequivalence and interchangeability between generics make sense only if the condition that both generics are bioequivalent to the corresponding brand name drug holds. However, directly adopting adjusted indirect comparison does not take the condition into consideration. In view of this, we proposed the restricted confidence interval of logarithmic geometric mean ratio (GMR) for two generics, under the framework of indirect comparisons with constraints. Furthermore, the fiducial probability [15] of bioequivalence between generics can be derived. In this article, we focused on bioequivalence testing of a basket of generics and propose a normality-based method to screen average bioequivalence pairs of generics with certain assurance. Assume that there are two generics, denoted by G_A and G_B, respectively, which have been shown to be bioequivalent to the same brand name drug (denoted by B_R). The available data are from two bioequivalence trails, one comparing G_A and B_R and the other one comparing G_B and B_R. The same criterion used for testing the average bioequivalence of a generic and the brand name drug is adopted to address the testing of average bioequivalence between two generics. That is, if the 90% confidence interval of the GMR between two generics falls into a prespecified interval, say 80.00% and 125.00%, we claim that the two generics are average bioequivalent. The confidence interval of GMR is constructed based on the 90% confidence intervals from the original bioequivalence studies under normality assumption.

In the next section, we briefly discussed the adjusted indirect comparison for testing bioequivalence between generics, outlined the proposed methods, and provided detailed description under the conventional two-sequence, two-period crossover design. In Section 3, some extensions for the proposed methods including multiple comparisons were presented. Then Section 4 discussed the important assumptions of similarity and the determination of the bioequivalence limits. In Section 5, we conducted extensive simulation studies to investigate the performance of the proposed methods. Two examples of malaria generics and HIV/AIDS generics were provided to demonstrate the use of the method in Section 6. In the end, some concluding remarks were given in Section 7.

2. Restricted confidence intervals of geometric mean ratio between generics

In this section, we discuss the inappropriateness of directly applying adjusted indirect comparison to testing generics' bioequivalence and then construct two types of restricted confidence intervals of GMR between two generics, using the normality-based statistics from two independent bioequivalence trials. The applicability of the proposed methods depends on similarity assumptions, which will be discussed in Section 4. To better illustrate the method, we provided detailed solutions under 2×2 crossover design without carryover effect in the Appendix.

Denote (L_A, U_A) and (L_B, U_B) as the $1 - 2\alpha$ confidence intervals of $\mu_A - \mu_R$ and $\mu_B - \mu_R$, respectively, where μ_A, μ_B, and μ_R are the true logarithmic geometric means of G_A, G_B, and B_R, respectively. (L_A, U_A)

and (L_B, U_B) were obtained from two independent bioequivalence trials with normality-based statistical methods. Assume that the two trials are similar so that the relative effect estimated by the trial of G_A versus B_R is generalizable to subjects in the trial of G_B versus B_R and the relative effect estimated by the trial of G_B versus B_R is generalizable to subjects in the trial of G_A versus B_R. Denote $(\delta_L, \delta_U) = (\log(0.8), \log(1.25))$ as the bioequivalence limits. The approval of both generics require that both (L_A, U_A) and (L_B, U_B) fall within (δ_L, δ_U). Based on (L_A, U_A) and (L_B, U_B), the statistical methods used for obtaining these two confidence intervals, and the similarity assumptions, we construct the restricted confidence intervals of $\mu_A - \mu_B$ by indirect comparison.

2.1. Adjusted indirect comparison for generics' bioequivalence

The method of adjusted indirect comparison can be described as follows. In the trial of G_A versus B_R, assume that $\mu_A - \mu_R$ was estimated as $\hat{v}_{AR}$. Similarly, in the trial of G_B versus B_R, $\mu_B - \mu_R$ was estimated as $\hat{v}_{BR}$. Under the similarity assumption, based on indirect comparison, we get a point estimator of $\mu_A - \mu_B$ as $\hat{v}_{AR} - \hat{v}_{BR}$, of which the variance was estimated by $\widehat{Var}_{AB}$. Under the normality assumption, the $1 - 2\alpha$ confidence interval of $\mu_A - \mu_B$ for adjusted indirect comparison can be expressed as follows:

$$\hat{v}_{AR} - \hat{v}_{BR} \pm z_\alpha / t_\alpha(d.f.) \cdot \sqrt{\widehat{Var}_{AB}}, \tag{1}$$

where $z_\alpha / t_\alpha(d.f.)$ is the α quantile of standard normal distribution or that of the Student's t distribution with the degree of freedom $d.f.$

From the viewpoint of Gwaza *et al.* [13], among the available methods for performing indirect comparisons, the adjusted indirect comparison is the simplest and most suitable method for bioequivalence studies, because it uses publicly available data and partly preserves the power of randomized controlled trials. For clinical practice, the bioequivalence and interchangeability between generics make sense only if both generics are bioequivalent to the corresponding brand name drug. Directly adopting adjusted indirect comparison may not be appropriate because this method derives the confidence interval of $\mu_A - \mu_B$ ignoring the fact that the confidence intervals for $\mu_A - \mu_R$ and $\mu_B - \mu_R$ are contained within (δ_L, δ_U), resulting in a narrower confidence interval and overestimation of clinical meaningful bioequivalence between generics. In order to avoid this limitation, we imposed constraints on the indirect comparison and proposed two types of restricted confidence intervals of $\mu_A - \mu_B$.

2.2. Restricted confidence intervals of $\mu_A - \mu_B$

In this subsection, we present two types of restricted confidence intervals of $\mu_A - \mu_B$. One is based on indirect comparisons under the constraint of bioequivalence between the generics and the brand name drug. The other is a continuous version of the first one.

2.2.1. Type 1 restricted confidence interval. Without loss of generality, in the trial of G_A versus B_R, assume that $\mu_A - \mu_R$ was estimated as $\hat{v}_{AR}$, of which the variance was estimated by $\widehat{Var}(\hat{v}_{AR})$. (L_A, U_A) can be expressed as $\hat{v}_{AR} \pm t_\alpha(df_A)\sqrt{\widehat{Var}(\hat{v}_{AR})}$, where $t_\alpha(df_A)$ is the α quantile of the Student's t distribution with the degree of freedom df_A. Note that $t_\alpha(df_A)$ is the α quantile of standard normal distribution when $df_A = \infty$. Similarly, in the trial of G_B versus B_R, $\mu_B - \mu_R$ was estimated as $\hat{v}_{BR}$, of which the variance was estimated by $\widehat{Var}(\hat{v}_{BR})$, and $(L_B, U_B) = \hat{v}_{BR} \pm t_\alpha(df_B)\sqrt{\widehat{Var}(\hat{v}_{BR})}$. Then we obtain the fiducial distribution of $\mu_l - \mu_R$ as $\hat{v}_{lR} + t(df_l)\sqrt{\widehat{Var}(\hat{v}_{lR})}$, where $l = A\ or\ B$, and $t(df_l)$ is the Student's t distribution with the degree of freedom df_l. Denote f_l as the probability density function of the fiducial distribution of $\mu_l - \mu_R$. It is reasonable to assume that f_A and f_B are statistically independent. Based on f_A and f_B, calculate the following probability:

$$pr\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B\ and\ -\Delta \leqslant \mu_A - \mu_B \leqslant \Delta\},$$

denoted as q_Δ. When $\Delta = 2\delta_U$, $q_{2\delta_U}$ is equal to $pr\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B\}$, because $\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B\}$ is a subset of $\{2\delta_L \leqslant \mu_A - \mu_B \leqslant 2\delta_U\}$. For any q that $0 < q \leqslant q_{2\delta_U}$, we can find the minimal Δ satisfying

$$pr\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B\ and\ -\Delta \leqslant \mu_A - \mu_B \leqslant \Delta\} \geqslant q.$$

Denote this minimal Δ by Δ_q. Then we get a restricted confidence interval of $\mu_A - \mu_B$ as $(-\Delta_q, \Delta_q)$, with the confidence level of q. Denote $(\delta_L^{'}, \delta_U^{'})$ $(\delta_L^{'} = -\delta_U^{'})$ as the target bioequivalence limits for $\mu_A - \mu_B$. Given the confidence level of $1 - 2\beta$, if the resulting $\Delta_{1-2\beta} \leqslant \delta_U^{'}$, the clinical meaningful bioequivalence between G_A and G_B can be concluded. In other words, in this case, we have

$$pr\left\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B \text{ and } \delta_L^{'} \leqslant \mu_A - \mu_B \leqslant \delta_U^{'}\right\} \geqslant 1 - 2\beta.$$

Conversely, we can simply calculate the fiducial probability $pr\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B \text{ and } \delta_L^{'} \leqslant \mu_A - \mu_B \leqslant \delta_U^{'}\}$ (denoted as $q_{\delta_U^{'}}$) and then compare $q_{\delta_U^{'}}$ with the pre-specified confidence level. If $q_{\delta_U^{'}}$ is greater than the pre-specified confidence level, we conclude that the clinical meaningful bioequivalence between G_A and G_B holds.

The fiducial probability can be expressed as follows:

$$\int_{\delta_L}^{\delta_U} \int_{\delta_L \vee (x+\delta_L^{'})}^{\delta_U \wedge (x+\delta_U^{'})} f_A(x) f_B(y) dy dx, \tag{2}$$

and numerical integration can be used to get an accurate estimation.

2.2.2. Type 2 restricted confidence interval. In addition, we give a continuous extension of type 1 restricted confidence interval.

Given $(L_A, U_A) \subset (\delta_L, \delta_U)$ and $(L_B, U_B) \subset (\delta_L, \delta_U)$, now, we consider an optimization procedure. With a confidence level of $1 - 2\beta$, look for an interval of (l, u) that minimizes $|u - l|$ under the constrains of

$$\begin{aligned} &P\{\mu_A - \mu_R \in (l, u)\} \geqslant \sqrt{1 - 2\beta}, \\ &P\{\mu_B - \mu_R \in (l, u)\} \geqslant \sqrt{1 - 2\beta}, \\ &(l, u) \in (\delta_L, \delta_U). \end{aligned} \tag{3}$$

If such an interval exists, denote it by $(L_{1-2\beta}, U_{1-2\beta})$. If $U_{1-2\alpha} - L_{1-2\alpha} \leqslant \delta_U^{'}$, we have

$$P\{\mu_l - \mu_R \in (L_{1-2\alpha}, U_{1-2\alpha})\ l = A, B \text{ and } \mu_A - \mu_B \in (\delta_L^{'}, \delta_U^{'})\} \geqslant 1 - 2\alpha$$

and thus conclude that G_A and G_B are average bioequivalent in terms of geometric mean. The details for this part can be found in the Appendix.

3. Some extensions

In this section, we give some extensions of the proposed methods: getting p-value to accommodate multiple comparison; a variant with relaxed constraints; and a natural extension to the comparison of three generics.

3.1. Multiple comparisons

We may encounter the issue of multiple comparisons (or multiple testing) when conducting the pairwise comparisons. In multiple comparisons, usually, we adjust the significance level in some way, so that the probability of observing at least one significant result due to chance remains below the desired significance level. Specifically in practical applications, the family-wise error rate and false discovery rate (FDR) [16] are often considered and controlled with appropriate methods.

Family-wise error rate ensures that the probability of committing a single false rejection is bounded by α. It focuses on a 'family' of statistical tests and is appropriate when a single false positive in this family is a problem. The Bonferroni's correction is the most common way to achieve this, which sets the significance cut-off at α/m (here, m is the number of comparisons). The Bonferroni's correction tends to be too conservative, leading to a high rate of false negative or lower power, especially when the tests are correlated or m is large.

False discovery rate is defined as the proportion of actual false positive among all discoveries (significant results) [17]. The Benjamini–Hochberg procedure [16] is a powerful technique for controlling the FDR. It works by estimating some rejection region so that on average, $FDR < \alpha$, where α is a

pre-specified target value of FDR. This procedure can be briefly described as follows: Put all p-values in order, from smallest to largest; the smallest one has a rank of $i = 1$, and then the next smallest has $i = 2$, and so on; compare each individual p-value to its own critical value, $(i\alpha)/m$, where m is the total number of tests; denote the largest p-value satisfying $p < (i\alpha)/m$ by P_{max}; all p-values no larger than P_{max} are significant. Although independence of test statistics was assumed in Benjamini and Hochberg [16], addressing positive dependence by Benjamini and Yekutieli [18] essentially assured users that this simple procedure was safe to use in many situations arising in practice, and the modification to general dependence is often not needed [17]. Therefore, the Benjamini–Hochberg procedure is still expected to perform well in this article's case, which is under the pairwise comparisons setting, a specific situation of dependence. The pre-specified value of the FDR should be chosen cautiously. In bioequivalence studies for generics, missing out some pairs that are actually bioequivalent will not lead to serious consequences. On the other hand, the cost of getting a false positive (leading to safety/efficacy issues) can be huge. Thus, in the case of this article, controlling a low FDR (say 0.1) is recommended. Another approach proposed by Storey [19,20], the positive FDR, deals with FDR from a different philosophy and in an opposite way. It works by first fixing the significance region and then estimating the corresponding FDR α.

In the context of this article, we recommend using the Benjamini–Hochberg procedure for controlling the FDR. The target value of the FDR should be small. Sometimes a 'Benjamini–Hochberg-adjusted p-value' is used. The adjusted p-value for a test is either the raw p-value times m/i or the adjusted p-value for the next higher raw p-value, whichever is smaller. If the adjusted p-value is smaller than the target FDR, the test is significant. In the real examples of the following section, this Benjamini–Hochberg-adjusted p-value was used for controlling the FDR.

No matter which method is adopted, p-value in each single test is required for multiple testings. We give the expression of p-value for each type of the proposed methods as follows.

Type 1. From Section 2.2.1, the p-value can be expressed as follows:

$$1 - q_{\delta_U'} = 1 - pr\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A, B \text{ and } \delta_L' \leqslant \mu_A - \mu_B \leqslant \delta_U'\}.$$

Type 2. We propose a procedure to get a pseudo p-value for the proposed method in Section 2. Look for a value of p that maximizes $1 - p$ under constraints of

$$\begin{aligned} &P\{\mu_A - \mu_R \in (l, u)\} \geqslant \sqrt{1-p}, \\ &P\{\mu_B - \mu_R \in (l, u)\} \geqslant \sqrt{1-p}, \\ &(l, u) \subset (\delta_L, \delta_U), \\ &u - l = \delta_U', \\ &\sqrt{1-p} \geqslant 1 - 2\alpha. \end{aligned} \tag{4}$$

If such a p exists, denote it as $\tilde{p}$. Then we have $P\{\mu_l - \mu_R \in (\delta_L, \delta_U),\ l = A, B,\ \text{and}\ \mu_A - \mu_B \in (\delta_L', \delta_U')\} \geqslant 1 - \tilde{p}$, and $\tilde{p}$ can be used as p-value for the test.

To accommodate the variant of relaxed constraints, the constraints of (4) can be modified as follows:

$$\begin{aligned} &P\{\mu_A - \mu_R \in (l, u)\} \geqslant \frac{1-p}{q}, \\ &P\{\mu_B - \mu_R \in (l, u)\} \geqslant q, \\ &(l, u) \subset (\delta_L, \delta_U), \\ &u - l = \delta_U', \\ &min\left\{\frac{1-p}{q}, q\right\} \geqslant 1 - 2\alpha. \end{aligned} \tag{5}$$

3.2. *Other extensions*

We also considered the relaxed constraints for type 2 method and the extension to accommodate the comparison of more than two generics simultaneously. The technical details of these extensions can be found in the Appendix.

4. Similarity assumptions and determining the bioequivalence limits

4.1. Similarity assumption

The methods proposed should be applied cautiously. Both clinical similarity and methodological similarity should be considered when using the methods of restricted confidence interval based on indirect comparison [11]. Firstly, the important conditions of the trials that could bias the estimated effect should be comparable. Patients should be similar enough in the bioequivalence trials under comparison, such that the relative effect estimated by the trial of G_A versus B_R is generalizable to patients in the trial of G_B versus B_R, and vice versa. For pharmacokinetic/pharmacodynamic (PK/PD) bioequivalence research, patient characteristics, the mode of drug administration, and parameter measurement should be alike between the trails involved. Secondly, methodological quality of the trials should be sufficiently similar. If the trials for comparison are similarly biased, it is easy to mathematically prove that the results of indirect comparison are unbiased [12]. In addition, trials with different designs might not be comparable as the formulation effect might not be expected to be the same between them. However, Garcia-Arieta *et al.* [12] considered results from conventional 2×2 crossover designs and replicate designs as combinable.

Compared with indirect comparison of efficiency trials, confidence in the clinical similarity and methodological quality of bioequivalence studies is often ensured because the basic designs of such studies are generally consistent [12, 14]. For example, in bioequivalence trials, the participants' characteristics are frequently alike that healthy adult volunteers within the age of 18–55 years are commonly defined. Besides, randomization of subjects in the allocation of sequence is always guaranteed.

4.2. Determining the bioequivalence limits

Unlike the bioequivalence trials between generics and the brand name drug that are well designed with sufficient power, indirect comparison between generics often has larger variability and reduced precision, leading to a low level of power where the bioequivalence limits are set to be the same as in the bioequivalence trial (say $(\log(0.80), \log(1.25))$). In order to enlarge the power, as recommended by Garcia-Arieta *et al.* [12, 13], a slightly wider bioequivalence interval (say $(\log(0.70), \log(1.00/0.70))$) may be used for indirect comparison. However, the simulation studies in the following section showed that such wider limits may not be needed for the type 1 method.

5. Simulation study

We conduct simulation studies to investigate the performances of the proposed methods described in Section 2. Assume that all bioequivalence trials were conducted under 2×2 crossover designs without carryover effects. In each bioequivalence trial between a generic and the brand name drug, 24 (or 48) subjects were simulated with equal number of subjects in each arm (i.e., treatment arm and reference arm). A variety of scenarios were considered with parameter specifications given in Table I. Each pair of two trials, both of which were related to the same brand name drug, was taken indirectly to test average bioequivalence between the corresponding two generics. For each comparison of a generic and the brand name drug, 400 data sets were generated. Thus, 160,000 (400×400) comparisons arose for each scenario. Let $\delta_L = log(0.8)$ and $\delta_U = log(1.25)$ as recommended by FDA. The confidence level of each trial was set to be 90%, that is, $\alpha = 0.05$. We first obtained the confidence interval of GMR between a generic and the brand name drug for each trial and then implemented the proposed methods for testing the average bioequivalence of two generics in terms of GMR. For each scenario, the simulation results are summarized using the following notations.

GMR: the GMR between G_A and G_B before log-transformation.
PR_{both}: the rate of both trials (G_A *vs* B_R and G_B *vs* B_R) passing the average bioequivalence tests.
CPR_1: the conditional passing rate of the type 2 method under the condition of both trials passing the average bioequivalence tests, with $\delta_U^{'} = \delta_U$.
$\bar{l}$: the average length of $|U_{1-2\alpha} - L_{1-2\alpha}|$ if it exists.
CPR_2: the conditional passing rate of the type 1 method under the condition of both trials passing the average bioequivalence tests, with $\delta_U^{'} = \delta_U$.
CPR_3: the conditional passing rate of the type 1 method under the condition of both trials passing the average bioequivalence tests, with $\delta_U^{'} = log(10/7)$ suggested by Garcia-Arieta [12].

Table I. Parameter specification for simulation studies

E_A	95	110	105	105	90	90	90	90	85	85
E_B	95	110	100	95	95	100	105	110	110	115
E_R	100	100	100	100	100	100	100	100	100	100
Std_A	10	10	20	20	10	10	20	20		
Std_B	10	20	10	20	10	20	10	20		
Std_R	10	10	10	10	20	20	20	20		
$\frac{\sigma^2_{BT}}{\sigma^2_{TT}} = \frac{\sigma^2_{BR}}{\sigma^2_{TR}}$	3/4	1/4								
ρ	0.5									
n_A	24	24	48	48						
n_B	24	48	24	48						

1. E_A, E_B, and E_R are the means of generic G_A, generic G_B, and brand name drug B_R before log-transformation, respectively.
2. Std_A, Std_B, and Std_R are the total standard deviations of generic G_A, generic G_B, and brand name drug B_R before log-transformation, respectively.
3. σ^2_{BT}, σ^2_{TT}, σ^2_{BR}, σ^2_{TR}, ρ, n_A, and n_B have the same meanings as those in Appendix A.2.
4. The values of $\frac{\sigma^2_{BT}}{\sigma^2_{TT}}$ and the values of $\frac{\sigma^2_{BR}}{\sigma^2_{TR}}$ are set to be equal for the trial of G_A *vs* B_R and the trial of G_B *vs* B_R.
5. Without loss of generality, the fixed effects of periods and sequences in all bioequivalence trials are set to be 0 for all scenarios.
6. There are 10 combinations of (E_A, E_B, E_R), 8 combinations of (Std_A, Std_B, Std_R), and 4 combinations of (n_A, n_B), resulting in $10 \times 8 \times 4$ scenarios in total.

For simplicity, we only show the results for (Std_A, Std_B, Std_R) of $(20, 20, 20)$ with the value of $\frac{\sigma^2_{BT}}{\sigma^2_{TT}} = \frac{\sigma^2_{BR}}{\sigma^2_{TR}}$ as $3/4$ and $1/4$ in Tables II and III, respectively. The two tables represent two different levels of the intra-variation proportion. The results for other cases were similar, and we do not put them in this article. With Std_A, Std_B, Std_R, and $\sigma^2_{BT}/\sigma^2_{TT}$ ($\sigma^2_{BR}/\sigma^2_{TR}$) fixed, we have the following observations.

CPR_1: CPR_1 can be larger than 90% when GMR is very close to 1. When GMR departs from 1, with the other parameters fixed, CPR_1 gets smaller quickly. CPR_1 decreases to around 50% or even smaller with GMR of 0.9 or 1.1. When GMR is about 0.8, CPR_1 is very small (around 5% or even smaller). When sample size becomes larger, with the other parameters fixed, CPR_1 gets larger for log GMR within the bioequivalence limits but not close to δ_L (δ_U). Conversely, for log GMR close to δ_L (δ_U) or beyond the bioequivalence limits, with sample size getting larger, CPR_1 gets smaller.
$\bar{l}$: When GMR departs more from 1, with the other parameters fixed, $\bar{l}$ gets larger. When sample sizes become larger, with the other parameters fixed, $\bar{l}$ gets smaller.
CPR_2: CPR_2 are larger than 95% when GMR is very close to 1. When GMR departs from 1, with the other parameters fixed, the speed that CPR_2 gets smaller is much slower than that of CPR_1. CPR_2 is still around 80% or even larger with GMR of 0.9 or 1.1. When GMR is about 0.8, CPR_2 becomes small (around 20%). When sample size becomes larger, with the other parameters fixed, the trend of CPR_2 is similar to that of CPR_1.
CPR_3: CPR_3 are nearly 100% when GMR is very close to 1. When GMR departs from 1, CPR_3 decreases pretty slowly, with a value of larger than 95% for GMR of around 0.8.

With E_A, E_B, and E_R fixed, we have the following observations:

CPR_1: For log GMR within the bioequivalence limits but not close to δ_L (δ_U), when Std_A and/or Std_B become larger, with the other parameters fixed, CPR_1 gets smaller. When Std_R becomes larger, with the other parameters fixed, CPR_1 gets smaller. For log GMR beyond the bioequivalence limits, CPR_1 changes in the opposite direction compared with the aforementioned cases. If log GMR is within the bioequivalence limits but not close to δ_L (δ_U), when $\sigma^2_{BT}/\sigma^2_{TT}$ ($\sigma^2_{BR}/\sigma^2_{TR}$) becomes smaller, with the other parameters fixed, CPR_1 gets smaller. If log GMR is beyond the bioequivalence limits, CPR_1 changes in the opposite direction compared with the aforementioned case. The trends of CPR_2 and CPR_3 are similar to that of CPR_1.
$\bar{l}$: When Std_A and/or Std_B become larger, with the other parameters fixed, $\bar{l}$ gets larger. When Std_R becomes larger, with the other parameters fixed, $\bar{l}$ gets larger.

Table II. Simulation results: when $Std_A = 20$, $Std_B = 20$, and $Std_R = 20$ and when $\frac{\sigma^2_{BT}}{\sigma^2_{TT}} = \frac{\sigma^2_{BR}}{\sigma^2_{TR}} = 3/4$

E_A	E_B	Geometric mean ratio	n_A	n_B	PR_{both}	CPR_1	$\bar{l}$	CPR_2	CPR_3
95	95	1	24	24	0.912	0.437	0.232	0.969	1
			24	48	0.943	0.683	0.21	0.99	1
			48	24	0.962	0.718	0.21	0.993	1
			48	48	0.995	0.976	0.16	0.999	1
110	110	1	24	24	0.658	0.607	0.219	0.995	1
			24	48	0.792	0.819	0.203	0.998	1
			48	24	0.792	0.79	0.203	0.998	1
			48	48	0.953	0.996	0.156	1	1
105	100	1.052	24	24	0.958	0.446	0.236	0.909	1
			24	48	0.965	0.688	0.208	0.972	1
			48	24	0.992	0.649	0.214	0.956	1
			48	48	1	0.92	0.169	0.993	1
105	95	1.11	24	24	0.914	0.232	0.265	0.734	1
			24	48	0.965	0.385	0.243	0.795	1
			48	24	0.948	0.35	0.243	0.828	1
			48	48	1	0.579	0.215	0.903	1
90	95	0.945	24	24	0.63	0.328	0.24	0.952	1
			24	48	0.665	0.578	0.219	0.994	1
			48	24	0.85	0.519	0.225	0.942	1
			48	48	0.898	0.888	0.181	0.994	1
90	100	0.896	24	24	0.66	0.21	0.261	0.799	0.999
			24	48	0.665	0.378	0.235	0.925	1
			48	24	0.891	0.279	0.254	0.76	0.999
			48	48	0.898	0.536	0.22	0.907	1
90	105	0.852	24	24	0.642	0.103	0.295	0.511	0.996
			24	48	0.665	0.151	0.274	0.588	0.999
			48	24	0.866	0.103	0.293	0.443	0.994
			48	48	0.898	0.144	0.271	0.536	0.999
90	110	0.812	24	24	0.53	0.027	0.332	0.238	0.978
			24	48	0.638	0.032	0.321	0.224	0.989
			48	24	0.716	0.013	0.332	0.161	0.975
			48	48	0.862	0.013	0.321	0.146	0.985
85	110	0.765	24	24	0.167	0.004	0.353	0.118	0.951
			24	48	0.202	0.005	0.343	0.097	0.976
			48	24	0.233	0	0.362	0.033	0.919
			48	48	0.281	0	0.352	0.019	0.954
85	115	0.73	24	24	0.084	0.002	0.371	0.052	0.912
			24	48	0.135	0	0.376	0.018	0.862
			48	24	0.117	0	0.381	0.011	0.849
			48	48	0.189	0	0.386	0	0.773

From those observations, we see the proposed type 1 method using $\delta'_U = \delta_U$ performs well for the given scenarios. When the noises of the data are not large and the true log GMR is within the bioequivalence limits but not close to δ_L (δ_U), this method has large passing rate of the average bioequivalence testing between two generics in terms of GMR. When the true log GMR is beyond the bioequivalence limits, the passing rate is very small. While the adjusted indirect comparison puts a constraint that the point estimation ratios should not exceed the 7% difference to ensure interchangeability between generics, simulation results indicated that the proposed type I method can relax this constraint up to 15% if the original studies were sufficiently powered. Therefore, the type 1 method has larger power. Besides the scenarios presented here, we also considered many others and carried out similar simulation studies. The results are similar to those presented here, and we did not include them in this article for simplicity.

Table III. Simulation results: when $Std_A = 20$, $Std_B = 20$, and $Std_R = 20$ and when $\frac{\sigma^2_{BT}}{\sigma^2_{TT}} = \frac{\sigma^2_{BR}}{\sigma^2_{TR}} = 1/4$

E_A	E_B	Geometric mean ratio	n_A	n_B	PR_{both}	CPR_1	$\bar{l}$	CPR_2	CPR_3
95	95	1	24	24	0.77	0.082	0.269	0.912	1
			24	48	0.866	0.292	0.245	0.958	1
			48	24	0.876	0.274	0.246	0.967	1
			48	48	0.985	0.891	0.188	0.994	1
110	110	1	24	24	0.469	0.173	0.251	0.972	1
			24	48	0.616	0.397	0.235	0.981	1
			48	24	0.632	0.43	0.23	0.984	1
			48	48	0.83	0.951	0.18	0.999	1
105	100	1.052	24	24	0.877	0.112	0.269	0.849	1
			24	48	0.898	0.328	0.241	0.938	1
			48	24	0.97	0.312	0.245	0.91	1
			48	48	0.992	0.813	0.192	0.976	1
105	95	1.11	24	24	0.783	0.064	0.291	0.67	0.995
			24	48	0.891	0.161	0.271	0.708	0.999
			48	24	0.866	0.174	0.268	0.752	1
			48	48	0.985	0.421	0.237	0.808	1
90	95	0.945	24	24	0.441	0.054	0.273	0.894	1
			24	48	0.501	0.234	0.248	0.985	1
			48	24	0.713	0.194	0.258	0.88	0.999
			48	48	0.811	0.783	0.201	0.986	1
90	100	0.896	24	24	0.494	0.04	0.288	0.73	0.999
			24	48	0.505	0.128	0.264	0.852	1
			48	24	0.799	0.126	0.276	0.692	0.998
			48	48	0.818	0.392	0.239	0.825	1
90	105	0.852	24	24	0.453	0.018	0.317	0.46	0.989
			24	48	0.501	0.048	0.294	0.535	0.999
			48	24	0.734	0.048	0.31	0.411	0.985
			48	48	0.811	0.124	0.283	0.496	0.998
90	110	0.812	24	24	0.341	0.009	0.339	0.274	0.982
			24	48	0.46	0.012	0.333	0.222	0.977
			48	24	0.552	0.02	0.333	0.226	0.975
			48	48	0.744	0.019	0.327	0.185	0.973
85	110	0.765	24	24	0.098	0.004	0.348	0.204	0.983
			24	48	0.132	0.002	0.342	0.149	0.976
			48	24	0.196	0.006	0.356	0.115	0.934
			48	48	0.264	0.005	0.35	0.072	0.935
85	115	0.73	24	24	0.047	0	0.365	0.079	0.968
			24	48	0.088	0	0.37	0.027	0.933
			48	24	0.094	0.001	0.374	0.034	0.889
			48	48	0.177	0	0.38	0.012	0.822

6. Real example analysis

In this section, we used the type 1 method with $\delta'_U = \delta_U$ to investigate the relative bioavailability of different generics. Two examples were illustrated to demonstrate the use of the proposed method.

6.1. Malaria generics

As in Gwaza *et al.* [13], we selected the data from three bioavailability/bioequivalence studies conducted independently, which are available in WHO public assessment reports at https://extranet.who.int/prequal/ (https://extranet.who.int/prequal/sites/default/files/documents/MA052part6v1.pdf, https://extranet.who.int/prequal/sites/default/files/documents/MA062part6v1.pdf, and https://extranet.who.int/prequal/sites/default/files/documents/MA064part6v2.pdf). The studies compared fixed dose combination artemether/

Table IV. The 90% confidence interval of pharmacokinetic parameters of artemether and lumefantrine in fixed dose combination generics, the fiducial probability with $\delta_U' = \delta_U = \log(1.25)$, and the Benjamini–Hochberg-adjusted p-value

		The 90% confidence interval		
		Ajanta	Ipca	Cipla
Artemether	C_{max}	86.3–99.1	101.0–118.9	86.7–103.7
	AUC_{0-t}	93.3–104.7	102.4–116.6	90.1–106.9
	AUC_{0-inf}	93.5–104.6	102.8–117.0	90.1–106.9
Lumefantrine	C_{max}	80.3–95.4	87.6–99.5	89.3–108.2
	AUC_{0-t}	82.1–97.3	87.3–103.2	88.7–108.1
	AUC_{0-inf}	82.5–97.8	87.6–103.0	88.7–108.1
		The fiducial probability (the adjusted p-value)		
		Ajanta versus Ipca	Ajanta versus Cipla	Ipca versus Cipla
Artemether	C_{max}	0.795 (0.2048)*	0.996 (0.0238)	0.857 (0.1517)*
	AUC_{0-t}	0.989 (0.0255)	0.999 (0.0055)	0.962 (0.0623)
	AUC_{0-inf}	0.989 (0.0255)	0.999 (0.0055)	0.957 (0.0623)
Lumefantrine	C_{max}	0.955 (0.0623)	0.894 (0.1197)*	0.992 (0.0255)
	AUC_{0-t}	0.975 (0.0459)	0.943 (0.0688)	0.991 (0.0255)
	AUC_{0-inf}	0.980 (0.0405)	0.950 (0.0642)	0.992 (0.0255)

*Not significant by the type 1 method with the target false discovery rate of 0.1.

lumefantrine 20/120 mg tablets in adult healthy volunteers with the same brand name drug submitted to the WHO Prequalification of Medicines Programme. These generic tablets were all compared with the brand name drug, the WHO approved artemether/lumefantrine 20/120 mg tablets (Coartem® /Riamet®) from Novartis Pharma (Basel, Switzerland) and had been prequalified by the WHO. The three studies enrolled 55, 64, and 58 adult men, respectively, and conducted single-center, open-label, randomized, two-period, two-treatment, two-sequence, crossover studies under non-fasting conditions. Two types of measures, the area under the time concentration curve (AUC_{0-t} and AUC_{0-inf}) and the drug peak concentration (C_{max}), were investigated for bioequivalence between generics. In the study performed by Cipla, an AUC truncated at 72 h was reported. Therefore, three potential pairs of generics were tested for bioequivalence, resulting to 18 comparisons after multiplying the six measures (3/3 for artemether/lumefantrine). We adjusted the p-value to control the target FDR using the Benjamini–Hochberg procedure.

Table IV listed the 90% confidence intervals of pharmacokinetic parameters of artemether and lumefantrine from those reports. Based on those confidence intervals and the same size in each study, we calculated the fiducial probability for each pairwise comparison and the Benjamini–Hochberg-adjusted p-value. The results were shown in the same table. Using the proposed type 1 method with $\delta_U' = \delta_U = log(125.00\%)$ and the target FDR of 0.1, all comparisons were significant and thus considered as bioequivalent except for three pairs for C_{max}: Ajanta versus Ipca, Ipca versus Cipla of artemether, and Ajanta versus Cipla of lumefantrine. The results were consistent with those obtained by the adjusted indirect comparison method with the conventional 80–125% acceptance range in Gwaza *et al.* [13]. Although three pairs of comparisons were not significant, two of them would be significant if the target FDR was set to be 0.2. The non-significant results might be partially attributed to the reduced precision of the proposed indirect method.

6.2. *HIV/AIDS generics*

Combivir® (lamivudine/zidovudine) 150 mg/300 mg tablet is an antiviral medication containing a combination of lamivudine and zidovudine. It belongs to reverse transcriptase inhibitors and helps keep the HIV virus from reproducing in human body. Lamivudine/zidovudine 150 mg/300 mg tablets are indicated for the treatment of HIV-1 infection in combination with at least one other antiretroviral agent. Several bioequivalence studies have been conducted to compare generic lamivudine/zidovudine 150 mg/300 mg tablet with the brand name drug Combivir® (lamivudine/zidovudine) 150 mg/300 mg tablet. We selected three studies that were sufficiently similar to investigate the bioequivalence between generics, using the same method as in the previous example. The data were from WHO public assessment reports available at https://extranet.who.int/prequal/ (https://extranet.who.int/prequal/sites/default/files/documents/HA286part6v2.pdf, https://extranet.who.int/prequal/sites/default/files/documents/HA291part6v1.pdf, and

Table V. The 90% confidence interval of pharmacokinetic parameters of lamivudine and zidovudine in fixed dose combination generics, the fiducial probability with $\delta_U' = \delta_U = \log(1.25)$, and the Benjamini–Hochberg-adjusted p-value

		The 90% confidence interval		
		Ranbaxy	Strides	Matrix
Lamivudine	C_{max}	96.8–109	90.63–110.52	82.4–101.7
	AUC_{0-t}	97.8–105	96.31–108.29	88.0–100.6
	AUC_{0-inf}	98.2–105	96.42–107.92	88.8–100.7
Zidovudine	C_{max}	85.5–108	81.15–112.51	83.8–114.1
	AUC_{0-t}	93.7–102	92.77–105.41	97.3–109.7
	AUC_{0-inf}	93.7–102	92.63–105.15	97.4–109.5
		The fiducial probability (the adjusted p-value)		
		Ranbaxy versus Strides	Ranbaxy versus Matrix	Strides versus Matrix
Lamivudine	C_{max}	0.996 (0.0057)	0.928 (0.0834)	0.929 (0.0834)
	AUC_{0-t}	1.000 (0.0000)	0.999 (0.0016)	0.995 (0.0072)
	AUC_{0-inf}	1.000 (0.0000)	1.000 (0.0010)	0.997 (0.0041)
Zidovudine	C_{max}	0.909 (0.0966)	0.926 (0.0834)	0.866 (0.1344)*
	AUC_{0-t}	1.000 (0.0001)	1.000 (0.0004)	0.999 (0.0012)
	AUC_{0-inf}	1.000 (0.0001)	1.000 (0.0004)	0.999 (0.0012)

*Not significant by the type 1 method with the target false discovery rate of 0.1.

https://extranet.who.int/prequal/sites/default/files/documents/HA392part6v1.pdf). Each study was randomized, open label, two-treatment, two-period, two-sequence, single-dose, and crossover. They were all conducted in healthy human adult subjects under fasting conditions performed between 2005 and 2007. The number of subjects that completed the study and was utilized for analysis to establish pharmacokinetic parameters and to assess bioequivalence was 62, 31, and 43, respectively. Three bioavailability characteristics C_{max}, AUC_{0-t}, and AUC_{0-inf} for bioequivalence evaluation were taken.

Table V summarized the results and the 90% confidence intervals from the reports. All comparisons were significant and thus were taken as bioequivalent except for the pair of strides versus matrix for C_{max} of zidovudine, with an adjusted p-value of 0.1344. The adjusted p-value of that pair was slightly larger than the target value of 0.1, which may be partially because of the insufficient sample sizes of the original studies. The results indicated that those WHO prequalified lamivudine/zidovudine 150 mg/300 mg products are not only bioequivalent with the brand name drug Combivir® in terms of AUC and C_{max} but also bioequivalent between themselves with respect to AUC and C_{max}.

7. Summary

The article discussed the issue of bioequivalence and interchangeability between generics that had been shown to be bioequivalent to the same brand name drug, based on average bioequivalence in terms of GMR. Indirect comparison is a valuable way to address it. However, the recommended adjusted indirect comparison may be inappropriate from the aspect of clinical meaningful values for the parameters. The methods with a constraint for testing bioequivalence based on indirect comparison were proposed, taking the clinical meaningful value into account. Among them, the type 1 method with the equivalence limits identical to that of conventional bioequivalence trials was emphasized based on its simulation performances, which showed that this method can screen out generic pairs with adequate power. The passing rate (power) is high when the true GMR is close to 1 and data noises are not large. The passing rate decreases when the true GMR departs away from 1 and approaches the bioequivalence limits. With a moderate GMR, for example, 0.9, the method has a satisfactory power when the original studies were adequately powered. The methods can be extended to simultaneous comparison of three generics and

multiple testings. The applicability of the proposed methods requires the similarity assumptions, which should be cautiously examined in advance.

However, there are still limitations of the proposed methods and some issues remain unsettled. When the GMR is far away from 1 but still within the bioequivalence limits, or the GMR is moderate but the original trials were not sufficiently powered, this approach does not perform well because of limited power. More accurate and adaptive methods can be explored. Besides, to use this method, some caution should be exercised with regard to the similarity assumptions. Moreover, we only focus on average bioequivalence based on GMR in this article. In fact, population bioequivalence and individual bioequivalence deserve further investigation regarding the bioequivalence and interchangeability between generics of the same brand name drug.

Appendix A

A.1. Type 2 restricted confidence interval

We give a continuous extension of type 1 restricted confidence interval. Firstly, we consider a simple case: Based on $P\{\mu_A-\mu_R\in(L_A,U_A)\}=1-2\alpha$ and $P\{\mu_B-\mu_R\in(L_B,U_B)\}=1-2\alpha$, if $max\{U_A,U_B\}-min\{L_A,L_B\}\leqslant\delta_U^{'}$, we have

$$\begin{aligned}
&P\{\mu_A-\mu_B\in(\delta_L^{'},\delta_U^{'}),\mu_A-\mu_R\in(L_A,U_A),\mu_B-\mu_R\in(L_B,U_B)\}\\
\geqslant\ &P\{min\{L_A,L_B\}\leqslant\mu_A-\mu_R\leqslant max\{U_A,U_B\},\\
&\quad min\{L_A,L_B\}\leqslant\mu_B-\mu_R\leqslant max\{U_A,U_B\},\\
&\quad \mu_A-\mu_R\in(L_A,U_A),\mu_B-\mu_R\in(L_B,U_B)\}\\
=\ &P\{\mu_A-\mu_R\in(L_A,U_A),\mu_B-\mu_R\in(L_B,U_B)\}\\
\geqslant\ &1-4\alpha.
\end{aligned}$$

Here, $(min\{L_A,L_B\},max\{U_A,U_B\})$ is a restricted confidence interval of $\mu_A-\mu_B$ with the confidence level of at least $1-4\alpha$.

In the aforementioned scenario, it is likely that there exists such an interval (L_{AB},U_{AB}), which satisfies

$$\begin{aligned}
&(L_{AB},U_{AB})\subset(\delta_L.\delta_U),\\
&U_{AB}-L_{AB}\leqslant max\{U_A,U_B\}-min\{L_A,L_B\},\\
&P\{\mu_A-\mu_R\in(L_{AB},U_{AB})\}\geqslant\sqrt{1-2\alpha},\ and\\
&P\{\mu_B-\mu_R\in(L_{AB},U_{AB})\}\geqslant\sqrt{1-2\alpha}.
\end{aligned}$$

As the two bioequivalence trials are assumed independent, it is reasonable to assume that (L_A,U_A) and (L_B,U_B) are statistically independent. Assume $P\{\mu_A-\mu_R\in(L_{AB},U_{AB})\}\geqslant\sqrt{1-2\alpha}$ can be derived from (L_A,U_A) and $P\{\mu_B-\mu_R\in(L_{AB},U_{AB})\}\geqslant\sqrt{1-2\alpha}$ from (L_B,U_B), $\{\mu_A-\mu_R\in(L_{AB},U_{AB})\}$ and $\{\mu_B-\mu_R\in(L_{AB},U_{AB})\}$ are approximately statistically independent. If $U_{AB}-L_{AB}\leqslant\delta_U^{'}$, we have

$$\begin{aligned}
&P\{\mu_l-\mu_R\in(L_{AB},U_{AB})\ l=A,B\}\\
=&P\{\mu_l-\mu_R\in(L_{AB},U_{AB})\ l=A,B\ and\ \mu_A-\mu_B\in(\delta_L^{'},\delta_U^{'})\}\geqslant 1-2\alpha,
\end{aligned}$$

and the clinical meaningful average bioequivalence between G_A and G_B is established.

Given $(L_A,U_A)\subset(\delta_L,\delta_U)$ and $(L_B,U_B)\subset(\delta_L,\delta_U)$, now, we consider an optimization procedure. With a confidence level of $1-2\beta$, look for an interval of (l,u) that minimizes $|u-l|$ under the constrains of

$$\begin{aligned}
&P\{\mu_A-\mu_R\in(l,u)\}\geqslant\sqrt{1-2\beta},\\
&P\{\mu_B-\mu_R\in(l,u)\}\geqslant\sqrt{1-2\beta},\\
&(l,u)\in(\delta_L,\delta_U).
\end{aligned}$$

If such an interval exists, denote it by $(L_{1-2\beta},U_{1-2\beta})$.

If $U_{1-2\alpha} - L_{1-2\alpha} \leqslant \delta_U^{'}$, we have

$$P\{\mu_l - \mu_R \in (L_{1-2\alpha}, U_{1-2\alpha})\ l = A, B \text{ and } \mu_A - \mu_B \in (\delta_L^{'}, \delta_U^{'})\} \geqslant 1 - 2\alpha$$

and conclude that G_A and G_B are average bioequivalent in terms of geometric mean. In the following subsection, without loss of generality, we assume that the two bioequivalence trials are both 2×2 crossover designs without carryover effects and then construct the restricted confidence intervals of $\mu_A - \mu_B$ in detail.

A.2. Under 2 × 2 crossover design without carryover effects

Let Y_{ijk} be the log-transformation of the pharmacokinetic response of interest, of the ith subject in the jth period, and of the kth sequence of the trial. As suggested by the FDA, the following statistical model is useful in describing Y_{ijk}:

$$Y_{ijk} = \mu + F_l + P_j + Q_k + S_{ikl} + e_{ijk},$$

where μ is the overall mean; P_j is the fixed effect of the jth period, where $j = 1, 2$ and $\sum_{j=1}^{2} P_j = 0$; Q_k is the fixed effect of the kth sequence, where $k = 1, 2$, and $\sum_{k=1}^{2} Q_k = 0$; F_l is the direct fixed effect of the lth drug formulation ($F_T + F_R = 0$) when $j = k$, $l = T$, the test formulation; otherwise, $l = R$, the reference (brand name) formulation; S_{ikl} is the random effect of the ith subject in the kth sequence under drug formulation l; and $S_{ik} = (S_{ikT}, S_{ikR})$, $i = 1, \ldots, n_k$, and $k = 1, 2$ are independently and identically distributed bivariate normal random vectors with mean $(0, 0)$ and an unknown variance–covariance matrix:

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix};$$

e_{ijk}'s are the independent random errors distributed as $N(0, \sigma_{Wl}^2)$, and S_{ik}'s and e_{ijk}'s are independent. Note that σ_{BT}^2 and σ_{BR}^2 are between-subject variances and σ_{WT}^2 and σ_{WR}^2 are within-subject variances and that $\sigma_{TT}^2 = \sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_{TR}^2 = \sigma_{BR}^2 + \sigma_{WR}^2$ are called total variances for the test and reference formulations, respectively.

The average bioequivalence index is $\nu = F_T - F_R = \mu_T - \mu_R$. Let $\bar{y}_{jk}$ be the sample average of the observations in the jth period and the kth sequence. Under the assumed statistical model, $\bar{y}_{11} - \bar{y}_{21} \sim N(\nu + P_1 - P_2, \frac{\tau^2}{n_1})$ and $\bar{y}_{12} - \bar{y}_{22} \sim N(-\nu + P_1 - P_2, \frac{\tau^2}{n_2})$, where $\tau^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} + \sigma_{WT}^2 + \sigma_{WR}^2 = \sigma_{TT}^2 + \sigma_{TR}^2 - 2\rho\sigma_{BT}\sigma_{BR}$. Consequently, we have the point estimators of ν and τ^2 as $\hat{\nu} = \frac{1}{2}(\bar{y}_{11} - \bar{y}_{21} - \bar{y}_{12} + \bar{y}_{22})$ and $\hat{\tau}^2 = \frac{(n_1-1)s_{D1}^2 + (n_2-1)s_{D2}^2}{n_1+n_2-2}$, where s_{Dk}^2 is the sample variance based on the differences $\{y_{i1k} - y_{i2k}, i = 1, \ldots, n_k\}, k = 1, 2$.

Denote $c = \frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})$; the point estimators satisfy $\hat{\nu} \sim N(\nu, c\tau^2)$ and $\frac{(n_1+n_2-2)\hat{\tau}^2}{\tau^2} \sim \chi^2(n_1+n_2-2)$. Then, the $1-2\alpha$ confidence lower limit and upper limit of ν are given as $L = \hat{\nu} - Z_\alpha\sqrt{c\hat{\tau}^2}$ and $U = \hat{\nu} + Z_\alpha\sqrt{c\hat{\tau}^2}$, where $Z_\alpha = t_\alpha(n_1 + n_2 - 2)$ is the α quantile of t distribution with degree of $n_1 + n_2 - 2$.

Now, assume that both bioequivalence trials of G_A *vs* B_R and G_B *vs* B_R are 2 × 2 crossover designs without carryover effects. The sample sizes of the two trials were n_A and n_B, respectively. (L_A, U_A) and (L_B, U_B) are the $1 - 2\alpha$ confidence intervals of $\mu_A - \mu_R$ and $\mu_B - \mu_R$, respectively. Denote the estimators of (ν, τ) by $(\hat{\nu}_A, \hat{\tau}_A)$ for G_A and $(\hat{\nu}_B, \hat{\tau}_B)$ for G_B. Thus, we have

$$\begin{aligned} \hat{\nu}_A &= (L_A + U_A)/2, \\ \sqrt{c_A\hat{\tau}_A^2} &= \frac{U_A - L_A}{-2 \times t_\alpha(n_A - 2)}, \\ \hat{\nu}_B &= (L_B + U_B)/2, \\ \sqrt{c_B\hat{\tau}_B^2} &= \frac{U_B - L_B}{-2 \times t_\alpha(n_B - 2)}. \end{aligned}$$

Denote $\sqrt{c_A\hat{\tau}_A^2}$ and $\sqrt{c_B\hat{\tau}_B^2}$ by S_A and S_B, respectively.

Type 1. To test the clinical meaningful average bioequivalence between G_A and G_B based on type 1 restricted confidence interval, firstly, we obtain the fiducial distributions of $\mu_A - \mu_R$ and $\mu_B - \mu_R$. Taking $(\hat{\nu}_A, \hat{\nu}_B, S_A, S_B)$ as constants, we have $\mu_A - \mu_R \sim \hat{\nu}_A + t(n_A - 2)S_A$ and $\mu_B - \mu_R \sim \hat{\nu}_B + t(n_B - 2)S_B$, where

$t(n_A-2)$ and $t(n_B-2)$ are the Student's t distribution with the degree of freedom n_A-2 and n_B-2, respectively.

The fiducial probability $q_{\delta_U'} = pr\{\delta_L \leqslant \mu_l - \mu_R \leqslant \delta_U,\ l = A,B \text{ and } \delta_L' \leqslant \mu_A - \mu_B \leqslant \delta_U'\}$ can be expressed as follows:

$$\begin{aligned}
&\int_{\delta_L}^{\delta_U}\int_{\delta_L\vee(x+\delta_L')}^{\delta_U\wedge(x+\delta_U')} \frac{1}{S_A S_B} f_A\left(\frac{x-\hat{v}_A}{S_A}\right) f_B\left(\frac{y-\hat{v}_B}{S_B}\right) dydx \\
&= \int_{\delta_L}^{\delta_U} \frac{1}{S_A} f_A\left(\frac{x-\hat{v}_A}{S_A}\right) \\
&\qquad \cdot \left[F_B\left(\frac{(\delta_U\wedge(x+\delta_U'))\vee\delta_L - \hat{v}_B}{S_B}\right) - F_B\left(\frac{(\delta_L\vee(x+\delta_L'))\wedge\delta_U - \hat{v}_B}{S_B}\right)\right] dx,
\end{aligned}$$

where f_A and f_B are the probability density functions of $t(n_A-2)$ and $t(n_B-2)$, respectively, and F_B is the cumulative distribution function of $t(n_B-2)$. We can obtain $q_{\delta_U'}$ by numerical integration. If $q_{\delta_U'}$ is greater than the pre-specified confidence level, we conclude that generic G_A and generic G_B are average bioequivalent in terms of geometric mean.

Type 2. To test the clinical meaningful average bioequivalence between G_A and G_B based on type 2 restricted confidence interval, firstly, we check whether $(L_{1-2\alpha}, U_{1-2\alpha})$ exists: Calculate l_A' and l_B' such that $P\{\mu_A-\mu_R \in (l_A', \delta_U)\} = \sqrt{1-2\alpha}$ and $P\{\mu_B-\mu_R \in (l_B', \delta_U)\} = \sqrt{1-2\alpha}$; let $l' = min\{l_A', l_B'\}$; if $l' \geqslant \delta_L$, then $(L_{1-2\alpha}, U_{1-2\alpha})$ exists.

Here, we only show how to calculate l_A', and it is similar for l_B': Denote $l_A' = \hat{v}_A + t_\phi(n_A-2)S_A$ with ϕ unknown; calculate ϕ such that $\delta_U = \hat{v}_A + t_{\phi+\sqrt{1-2\alpha}}(n_A-2)S_A$; substitute the value of ϕ in $\hat{v}_A + t_\phi(n_A-2)S_A$ to get l_A'.

If $(L_{1-2\alpha}, U_{1-2\alpha})$ exists, we get

$$\begin{aligned}
L_{1-2\alpha} = argmin_l\{max\{u_A,u_B\} - l :\ & \\
& P\{\mu_A-\mu_R \in (l,u_A)\} = \sqrt{1-2\alpha}, \\
& P\{\mu_B-\mu_R \in (l,u_B)\} = \sqrt{1-2\alpha}, \\
& l \in (\delta_L, l')\}
\end{aligned}$$

and $U_{1-2\alpha} = max\{u_A,u_B\}$ such that $P\{\mu_A-\mu_R \in (L_{1-2\alpha},u_A)\} = \sqrt{1-2\alpha}$ and $P\{\mu_B-\mu_R \in (L_{1-2\alpha},u_B)\} = \sqrt{1-2\alpha}$. If $(L_{1-2\alpha}, U_{1-2\alpha})$ exists and $U_{1-2\alpha} - L_{1-2\alpha} \leqslant \delta_U'$, we have $P\{\mu_l-\mu_R \in (L_{1-2\alpha}, U_{1-2\alpha})\ l = A,B$ and $\mu_A-\mu_B \in (\delta_L', \delta_U')\} \geqslant 1-2\alpha$ and conclude that G_A and G_B are average bioequivalent in terms of geometric mean.

A.3. Other extensions

A.3.1. A variant with relaxed constraints. This extension only applies to the method of type 2 restricted confidence interval.

Type 2. The constraints of (3) required both probabilities no less than $\sqrt{1-2\beta}$. We can relax the constraints as follows:

$$\begin{aligned}
& P\{\mu_A-\mu_R \in (l,u)\} \geqslant \frac{1-2\beta}{q}, \\
& P\{\mu_B-\mu_R \in (l,u)\} \geqslant q, \\
& (l,u) \in (\delta_L, \delta_U), \\
& 1-2\beta \leqslant q \leqslant 1.
\end{aligned}$$

Then the length of $|U_{1-2\beta} - L_{1-2\beta}|$ is not longer but probably shorter than that obtained under constraints of (3).

A.3.2. Extend to comparisons of three generics. In practice, other than pairwise comparison, the comparison of a basket of generics as a whole may arise. Here, we consider testing average bioequivalence

of three generics as a whole in terms of geometric mean. Denote G_A, G_B, and G_C as the three generics corresponding to the same brand name drug B_R, with their logarithmic geometric means denoted by μ_A, μ_B, μ_C, and μ_R. Denote (L_A, U_A), (L_B, U_B), and (L_C, U_C) as three confidence intervals of logarithmic geometric means from three bioequivalence trials of G_A *vs* B_R, G_B *vs* B_R and G_C *vs* B_R, respectively.

Type 1. Denote f_l as the probability density function of the fiducial distribution of $\mu_l - \mu_R$. The fiducial probability $P\{\mu_l - \mu_R \in (\delta_L, \delta_U),\ l = A, B, C,\ \text{and}\ \mu_A - \mu_B,\ \mu_A - \mu_C, \mu_B - \mu_C \in (\delta_L^{'}, \delta_U^{'})\}$ (denoted as $q_{\delta_U^{'}}$) can be expressed as follows:

$$\int_{\delta_L}^{\delta_U} \int_{\delta_L \vee (x+\delta_L^{'})}^{\delta_U \wedge (x+\delta_U^{'})} \int_{\delta_L \vee (x+\delta_L^{'}) \vee (y+\delta_L^{'})}^{\delta_U \wedge (x+\delta_U^{'}) \wedge (y+\delta_U^{'})} f_A(x) f_B(y) f_C(z) dz dy dx.$$

Type 2. Given $(L_l, U_l) \subset (\delta_L, \delta_U), l = A, B, C$, consider the following optimization procedure: With a confidence level of $1 - 2\beta$, look for an interval of (l, u) that minimizes $|u - l|$ under the constrains of

$$\begin{aligned} &P\{\mu_A - \mu_R \in (l, u)\} \geqslant p_1, \\ &P\{\mu_B - \mu_R \in (l, u)\} \geqslant p_2, \\ &P\{\mu_C - \mu_R \in (l, u)\} \geqslant p_3, \\ &(l, u) \in (\delta_L, \delta_U), \end{aligned}$$

where

$$\begin{aligned} &p_1 \geqslant 1 - 2\beta, \\ &p_2 \geqslant 1 - 2\beta, \\ &p_3 \geqslant 1 - 2\beta, \\ &p_1 p_2 p_3 \geqslant 1 - 2\beta. \end{aligned}$$

If such an interval exists, denote it by $(L_{1-2\beta}, U_{1-2\beta})$. If $(L_{1-2\alpha}, U_{1-2\alpha})$ exists and $U_{1-2\alpha} - L_{1-2\alpha} \leqslant \delta_U^{'}$, we have

$$P\{\mu_l - \mu_R \in (\delta_L, \delta_U),\ l = A, B, C,\ \text{and}\ \mu_A - \mu_B,\ \mu_A - \mu_C, \mu_B - \mu_C \in (\delta_L^{'}, \delta_U^{'})\} \geqslant 1 - 2\alpha.$$

Disclaimer

This article has been reviewed by FDA and determined not to be consistent with the Agency's views or policies. It reflects only the views and opinions of the authors.

Acknowledgements

We are very grateful to the joint editor, associate editor, and one referee for their insightful comments and constructive suggestions that have greatly improved this paper.

References

1. Chow SC, Liu JP. *Design and Analysis of Bioavailability and Bioequivalence Studies* 3rd ed. CRC Press: Boca Raton, 2009.
2. FDA. *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*. Center for Drug Evaluation and Research, U.S. Food and Drug Administration: Rockville, MD, 2001.
3. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; **15**(6):657–680.
4. Chow SC, Endrenyi L, Lachenbruch PA, Mentre F. Scientific factors and current issues in biosimilar studies. *Journal of Biopharmaceutical Statistics* 2014; **24**:1138–1153.
5. Chow SC, Song FY, Chen M. Some thoughts on drug interchangeability. *Journal of Biopharmaceutical Statistics* 2016; **26**:178–186.
6. Anderson S, Hauck WW. The transitivity of bioequivalence testing: potential for drift. *International Journal of Clinical Pharmacology and Therapeutics* 1996; **34**(9):369–374.
7. Bialer M, Midha KK. Generic products of antiepileptic drugs (AEDs): a perspective on bioequivalence and interchangeability. *Epilepsia* 2010; **51**:941–950.
8. Privitera M. Generic antiepileptic drugs: current controversies and future directions. *Epilepsy Curr* 2008; **8**:113–117.

9. van Gelder T, Gabardi S. Methods, strengths, weaknesses, and limitations of bioequivalence tests with special regard to immunosuppressive drugs. *Transplant Int* 2013; **26**(8):771–777.
10. FDA. *Draft Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endponits for Drugs Submitted under an ANDA*. Center for Drug Evaluation and Research, U.S. Food and Drug Administration: Rockville, MD, 2013.
11. Song F. *What is indirect comparison?* February 2009.
12. Garca-Arieta A, Potthast H, Leufkens H, Welink J, Leufkens H, Stah M, García-Arieta A. Assessment of the interchangeability between generics. *Generics and Biosimilars Initiative Journal (GaBI Journal)* 2016; **5**(2):55–59.
13. Gwaza L, Gordon J, Welink J, Potthast H, Hansson H, Stahl M, García-Arieta A. Statistical approaches to indirectly compare bioequivalence between generics: a comparison of methodologies employing artemether/lumefantrine 20/120 mg tablets as prequalified by WHO. *European Journal of Clinical Pharmacology* 2012; **68**:1611–1618.
14. Endrenyi L, Tóthfalusi L. Adjusted indirect comparisons between generics – bioequivalence and interchangeability. *Generics and Biosimilars Initiative Journal (GaBI Journal)* 2016; **5**(2):53–54.
15. Fisher RA. The fiducial argument in statistical inference. *Annals of Human Genetics* 1935; **6**(4):391–398.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; **57**(1):289–300.
17. Benjamini Y. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010; **72**(4):405–416.
18. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001; **29**:1165–1188.
19. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(3):479–498.
20. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *PNAS, Proceedings of the National Academy of Sciences* 2003; **100**(16):9440–9445.