ATLANTIS PRESS Journal of Statistical Theory and Applications, Vol. 16, No. 4 (December 2017) 508–521

On the Assessment of Average Biosimilarity Based on a Three-Arm Parallel Design

Ginto Jacob Pottackal

Department of Mathematics and Statistics
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, Maryland 21250
ginto1@umbc.edu

Thomas Mathew

Department of Mathematics and Statistics
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, Maryland 21250
mathew@umbc.edu

Received 23 November 2016

Accepted 30 May 2017

Average biosimilarity is investigated under a three-arm parallel design: one arm corresponds to the test drug T, and the other two arms correspond to two versions of the reference drug, say R_1 and R_2. The hypothesis of interest is the equivalence of the population average response for T with the mean of the population average responses for R_1 and R_2. The parameter of interest is formulated as the absolute difference of the above two averages, scaled by the absolute difference between the population means corresponding to R_1 and R_2. A difference parameter is also proposed. For the ratio parameter, a test can be derived using the asymptotic normality of an appropriate test statistic; however, the test is not satisfactory in terms of type I error probabilities. Improved tests are derived by applying a bootstrap calibration, and by using the idea of a generalized pivotal quantity (gpq). The tests are developed under equal variance and unequal variance scenarios. Sample size determination is also addressed. For the difference parameter, a satisfactory test is developed using the gpq idea. The proposed methods result in tests that are satisfactory in terms of type I error performance

Keywords: Bootstrap calibration; equivalence testing; generalized pivotal quantity; sample size determination

2000 Mathematics Subject Classification: 22E46, 53C35, 57S20

1. Introduction

Biosimilar drugs are *highly similar* products or imitations of already approved biological drugs. Unlike generic drugs, it is difficult to develop exact copies of biological products due to the complexity of the protein structure. For the approval of generic drug products, the commonly used method is to assess average bioequivalence (ABE) regarding drug absorption through the conduct of bioequivalence studies. However, such a criterion alone may not be appropriate for concluding biosimilarity; nevertheless, the equivalence of averages should be a minimum requirement. We refer to the U.S. FDA guidance document [6] and the book by Chow [2] for further background information on biosimilars. In particular, Chow's book provides a thorough discussion, including a discussion of the various statistical criteria that can be used for establishing biosimilarity.

The pharmacokinetic (PK) response that is typically used for assessing bioequivalence or biosimilarity consists of the area under the blood or plasma concentration time curve (AUC). Furthermore, for generating the data for the assessment of biosimilarity, parallel study designs are found to be more practical. Indeed, the FDA guidance document [6] recommends a parallel design for establishing biosimilarity. Since we are comparing two drugs, the original biotechnology drug (or reference drug, denoted by R), and a copy (the test drug, denoted by T), a two arm parallel design appears natural. Here the subjects are randomized between the two arms with one arm corresponding to the reference drug R, and the second arm corresponding to the test drug T. Since exact copies of biotechnology drugs are not possible, some authors suggest that R should be compared with itself, and such information should be used to assess the similarity between T and R; see the article by Kang and Chow [7]. These authors thus suggest a three-arm parallel design for assessing biosimilarity; one arm of the design corresponds to the test drug T, and other two arms correspond to the reference drug coming from two batches, or manufactured using two processes; we shall denote these two versions of the reference drug by R_1 and R_2. Subjects are randomized among the three arms. Three-arm trials for equivalence assessment are also discussed in Chang et al. [1]. However, the third arm considered by the authors is a placebo, unlike the three-arm trial considered in Kang and Chow [7].

The set up we shall use is the same as that in Kang and Chow [7]. Let μ_T, μ_{R_1} and μ_{R_2} denote the population mean responses corresponding to T, R_1 and R_2. The responses could be AUC or $C_{\max}$, very often after a log-transformation. In order to assess biosimilarity with respect to averages, the parameter suggested by Kang and Chow [7] is the ratio

$$\theta = (\mu_T - \mu_R)/(\mu_{R_1} - \mu_{R_2}), \text{ where } \mu_R = (\mu_{R_1} + \mu_{R_2})/2, \tag{1.1}$$

and the hypotheses of interest are

$$H_0 : |\theta| \geq \delta, \text{ vs } H_1 : |\theta| < \delta, \tag{1.2}$$

for some specified threshold δ. Biosimilarity is concluded with respect to averages if H_0 is rejected based on a statistical test. Here it is assumed that $\mu_{R_1} \neq \mu_{R_2}$, as in Kang and Chow [7]. The justification for choosing the parameter θ as defined above is that one cannot expect μ_T to be close to μ_R any more than the amount by which the population averages are close for two copies of R. However, there are other parameter choices that can capture this requirement; for example, we can consider the difference

$$\theta_1 = |\mu_T - \mu_R| - |\mu_{R_1} - \mu_{R_2}|, \tag{1.3}$$

where μ_R is as defined in (1.1). Now the hypotheses of interest are

$$H_0 : |\theta_1| \geq \delta_1, \text{ vs } H_1 : |\theta_1| < \delta_1, \tag{1.4}$$

for some specified threshold δ_1. Biosimilarity is concluded with respect to averages if H_0 in (1.4) is rejected.

The purpose of our investigation is to develop appropriate tests for the hypothesis in (1.2) and (1.4). We shall assume normally distributed responses, as in Kang and Chow [7]. These authors have derived a test for the hypotheses in (1.2) using the asymptotic normality of the natural estimator of $\theta = (\mu_T - \mu_R)/(\mu_{R_1} - \mu_{R_2})$, assuming a common variance for the responses from the test drug and the reference drug. Even under such a scenario, Kang and Chow's approximate test is not always

satisfactory. We improve and generalize Kang and Chow's approach in the following ways. In the equal variance scenario, we apply a *bootstrap calibration* to the test due to Kang and Chow [7]. This results in a test that exhibits better performance in terms of type I error probability, but is still not entirely satisfactory. We also developed a test using the concept of a generalized pivotal quantity (gpq). The gpq-based test turned out to be the most satisfactory in terms of maintaining the type I error probability. We have also provided results on the power. We also consider the unequal variance situation, and pursue the same approaches to arrive at test procedures. We have also provided a table of sample sizes. The relevant results appear in Section 2 of the paper. Following this, we consider the hypotheses in (1.4) concerning the parameter $\theta_1 = (\mu_T - \mu_R) - (\mu_{R_1} - \mu_{R_2})$, and have investigated a gpq-based test. Simulation results show that such a test is satisfactory in terms of type I error probabilities. Details appear in Section 3 of the paper. Overall, our work has resulted in very satisfactory test procedures for assessing biosimilarity with respect to averages based on the hypotheses in (1.2) and (1.4). An example is presented in Section 4 of the paper. The example deals with testing the biosimilarity of Accofil, a biologic that is expected to be similar to the reference product Neupogen; the latter being used to boost white blood cell production in patients undergoing chemotherapy for certain cancers. Here we have a three arm design since Accofil will be compared to two versions of Neupogen: a version approved by the European Union and another version approved by the US. The relevant data are taken from a European Medicines Agency document [5]. Our methodology will be illustrated using this data set.

2. Testing Average Biosimilarity Based on $\theta = (\mu_T - \mu_R)/(\mu_{R_1} - \mu_{R_2})$

Recall that our problem is that of testing the hypotheses in (1.2) based on data generated using a three-arm parallel design. Under such a design, let n_T subjects receive the test drug T, n_R subjects receive the reference drug R_1, and another group of n_R subjects receive the reference drug R_2, where R_1 and R_2 represent two versions of the reference drug R manufactured using two processes, or in two batches. In Kang and Chow [7], the authors assume a 2:1 ratio between n_T and n_R. We shall also assume this, even though this is not essential for the development of our methodology. It is assumed that the responses follow a normal distribution where the population means corresponding to the three arms are denoted by μ_T, μ_{R_1} and μ_{R_2}, respectively, as specified in the previous section. In their work, Kang and Chow [7] assume a common population variance σ^2 for the responses from T, as well as for the responses from R_1 and R_2. In our investigation, we shall first consider the case of a common variance, and later relax this assumption. In other words, we shall also consider the case of a population variance σ_T^2 for the responses for T, and a population variance σ_R^2 for the responses from R_1, as well as for those from R_2, where σ_T^2 and σ_R^2 need not be equal. We shall also denote by $\bar{X}_T$, $\bar{X}_{R_1}$ and $\bar{X}_{R_2}$, respectively, the sample means based on the responses from the sample of size n_T for T, the responses from the sample of size n_R for R_1, and those from the sample of size n_R for R_2. Furthermore, in the case of a common population variance σ^2, we shall denote by S^2 the unbiased estimator of σ^2 obtained by pooling the data from the three samples. We thus have the distributions

$$\bar{X}_T \sim N\left(\mu_T, \frac{\sigma^2}{n_T}\right),\ \bar{X}_{R_1} \sim N\left(\mu_{R_1}, \frac{\sigma^2}{n_R}\right),\ \bar{X}_{R_2} \sim N\left(\mu_{R_2}, \frac{\sigma^2}{n_R}\right),$$
$$(n_T + 2n_R - 3)\frac{S^2}{\sigma^2} \sim \chi^2_{n_T+2n_R-3}, \tag{2.1}$$

where χ^2_m denotes a central chisquare distribution with m df. Furthermore, the above random variables are independent. In the case of unequal population variances, let S_T^2 denote the sample variance from the sample of n_T responses for T, and let S_R^2 denote the sample variance obtained by pooling the n_R responses for R_1 and the n_R responses for R_2. We then have the distributions

$$\bar{X}_T \sim N\left(\mu_T, \frac{\sigma_T^2}{n_T}\right),\ \bar{X}_{R_1} \sim N\left(\mu_{R_1}, \frac{\sigma_R^2}{n_R}\right),\ \bar{X}_{R_2} \sim N\left(\mu_{R_2}, \frac{\sigma_R^2}{n_R}\right),$$
$$(n_T-1)\frac{S_T^2}{\sigma_T^2} \sim \chi^2_{n_T-1},\ (2n_R-2)\frac{S_R^2}{\sigma_R^2} \sim \chi^2_{2n_R-2}, \tag{2.2}$$

where the above random variables are also independent. In the common variance scenario, two test procedures are developed in Kang and Chow [7]: one based on the delta method, and a second test based on a linearization method. Based on numerical results, they then recommend the test based on the delta method, and also provide tables of sample sizes based on power considerations. In the common variance scenario, we shall first improve upon the delta method based test by applying a bootstrap calibration. Secondly, we shall extend the methodology to the unequal variance situation as well. In both cases, we shall also derive a test based on the idea of a generalized pivotal quantity. We shall also provide table of sample sizes so that our proposed tests will achieve a specified power.

2.1. *The generalized pivotal quantity (gpq)*

Before we describe the various test procedures, we shall briefly describe the generalized pivotal quantities that we shall use. The concept is due to Weerahandi [12]; see also [13] and [14]. A *generalized pivotal quantity* (gpq) is a function of the underlying random variables, and the observed data that are realizations of these random variables. We shall construct gpqs for μ_T, μ_{R_1} and μ_{R_2}, and then combine them to get a gpq for a parameter of interest (for example, the parameter θ in (1.1) and θ_1 in (1.3)). A gpq is required to satisfy two properties: (i) given the observed data, its distribution is free of any unknown parameters, and (ii) when the random variables are replaced by the corresponding realizations (i.e., the observed data), the gpq simplifies to a quantity that is free of any nuisance parameters; very often, the simplified quantity is equal to the parameter of interest.

In the equal variance scenario (2.1), let $\bar{x}_T, \bar{x}_{R_1}, \bar{x}_{R_2}$ and s^2 denote the observed values of $\bar{X}_T, \bar{X}_{R_1}$, $\bar{X}_{R_2}$ and S^2, respectively. Then gpqs for μ_T, μ_{R_1} and μ_{R_2}, denoted by $\tilde{\mu}_T$, $\tilde{\mu}_{R_1}$ and $\tilde{\mu}_{R_2}$, respectively, are given by

$$\begin{aligned}
\tilde{\mu}_T &= \bar{x}_T - \frac{Z_T}{U/\sqrt{n_T+2n_R-3}}\frac{s}{\sqrt{n_T}},\\
\tilde{\mu}_{R_1} &= \bar{x}_{R_1} + \frac{Z_{R_1}}{U/\sqrt{n_T+2n_R-3}}\frac{s}{\sqrt{n_R}},\\
\tilde{\mu}_{R_2} &= \bar{x}_{R_2} + \frac{Z_{R_2}}{U/\sqrt{n_T+2n_R-3}}\frac{s}{\sqrt{n_R}},
\end{aligned} \tag{2.3}$$

where Z_T, Z_{R_1} and Z_{R_2} are independent standard normal random variables and $U^2 \sim \chi^2$ with df = n_T+2n_R-3. We refer to the book by Krishnamoorthy and Mathew [9], Section 1.4, for further details on the derivation of the above gpqs. Having obtained the gpqs in (2.3), a gpq for θ, say $\tilde{\theta}$,

can be obtained as

$$\tilde{\theta} = \frac{|\tilde{\mu}_T - \tilde{\mu}_R|}{|\tilde{\mu}_{R_1} - \tilde{\mu}_{R_2}|}, \tag{2.4}$$

where $\tilde{\mu}_R = (\tilde{\mu}_{R_1} + \tilde{\mu}_{R_2})/2$. Percentiles of $\tilde{\theta}$ provide confidence limits for θ, referred to as *generalized confidence limits*.

In the unequal variance scenario (2.2), gpqs for μ_T, μ_{R_1} and μ_{R_2}, denoted once again by $\tilde{\mu}_T$, $\tilde{\mu}_{R_1}$ and $\tilde{\mu}_{R_2}$, respectively, are given by

$$\begin{aligned}
\tilde{\mu}_T &= \bar{x}_T - \frac{Z_T}{U_T/\sqrt{n_T - 1}} \frac{s_T}{\sqrt{n_T}}, \\
\tilde{\mu}_{R_1} &= \bar{x}_{R_1} + \frac{Z_{R_1}}{U_R/\sqrt{2n_R - 2}} \frac{s_R}{\sqrt{n_R}}, \\
\tilde{\mu}_{R_2} &= \bar{x}_{R_2} + \frac{Z_{R_2}}{U_R/\sqrt{2n_R - 2}} \frac{s_R}{\sqrt{n_R}},
\end{aligned} \tag{2.5}$$

where s_T^2 and s_R^2 are the observed values of S_T^2 and S_R^2, respectively, $U_T^2 \sim \chi^2$ with df = $n_T - 1$, $U_R^2 \sim \chi^2$ with df = $2n_R - 2$, and the rest of the quantities are as defined for the gpqs in (2.3).

2.2. *The case of a common variance*

We first give a brief description of the delta method approach given in Kang and Chow [7]. In view of the definition of θ in (1.1) and the distributions in (2.1), a natural estimator of θ, say $\hat{\theta}$, is given by

$$\hat{\theta} = \frac{\bar{X}_T - (\bar{X}_{R_1} + \bar{X}_{R_2})/2}{\bar{X}_{R_1} - \bar{X}_{R_2}}. \tag{2.6}$$

Defining

$$\mu_1 = \mu_T - (\mu_{R_1} + \mu_{R_2})/2,\ \mu_2 = \mu_{R_1} - \mu_{R_2},\ \sigma_1^2 = \frac{2\sigma^2}{n_T},\ \text{and}\ \sigma_2^2 = \frac{2\sigma^2}{n_R}, \tag{2.7}$$

$\hat{\theta}$ can also be expressed as

$$\hat{\theta} = \frac{\bar{V}}{\bar{U}},\ \text{ where } \bar{V} \sim N(\mu_1, \sigma_1^2) \text{ and } \bar{U} \sim N(\mu_2, \sigma_2^2), \tag{2.8}$$

where we have used the assumption $n_T = 2n_R$. We also note that $\bar{V}$ and $\bar{U}$ are also independent. A straightforward application of the delta method gives

$$\sqrt{n_T}\left(\frac{\bar{V}}{\bar{U}} - \frac{\mu_1}{\mu_2}\right) \sim N\left(0,\ \frac{2\sigma^2}{\mu_2^2} + \frac{4\mu_1^2\sigma^2}{\mu_2^4}\right), \tag{2.9}$$

asymptotically. Following the methodology used in bioequivalence testing, we reject H_0 in (1.2) when

$$Z = \frac{|\frac{\bar{V}}{\bar{U}}| - \delta}{\frac{S}{\sqrt{n_T}}\sqrt{\frac{2}{\bar{U}^2} + \frac{4\bar{V}^2}{\bar{U}^4}}} < -z_\alpha \tag{2.10}$$

where S^2 is the pooled variance mentioned in (2.1) and z_α is the upper α quantile of the standard normal distribution. Numerical results show that the performance of the above test is not satisfactory

in terms of maintaining the type I error probability; the type I error probabilities are often higher than the nominal level. We shall thus employ a bootstrap calibration in order to improve the performance.

2.2.1. *Bootstrap Calibration*

The bootstrap calibration idea that we shall apply is taken from Chapter 18 in the book by Efron and Tibshirani [4]. We shall explain the calibration idea as it applies to our problem. Since we noted that the test procedure in (2.10) has type I error probability more than the nominal level, it is possible that the type I error probability can be made closer to α by carrying out the test using a significance level $\gamma < \alpha$. The required significance level γ will be determined using the bootstrap. Under the normality assumption in (2.9), it is known that $\max_{H_0} P(Z < -z_\alpha) = P(Z < -z_\alpha | \theta = \delta)$, where H_0 is specified in (1.2). The above conclusion follows from the theoretical developments in the context of bioequivalence testing; see the book by Chow and Liu [3]. However, $P(Z < -z_\alpha | \theta = \delta)$ does depend on nuisance parameters. Since the type I error probability is being computed when $\theta = \mu_1/\mu_2 = \delta$, the nuisance parameters in the model can be taken as $\lambda = (\mu_1, \sigma^2)$, so that μ_2 is determined as $\mu_2 = \mu_1/\delta$. Thus, when $\theta = \mu_1/\mu_2 = \delta$, the distributions in (2.8) can be written as

$$\bar{V} \sim N\left(\mu_1, \frac{2\sigma^2}{n_T}\right) \text{ and } \bar{U} \sim N\left(\mu_1/\delta, \frac{2\sigma^2}{n_R}\right), \tag{2.11}$$

where we have used the expressions for σ_1^2 and σ_2^2 given in (2.7). We note that based on (2.11), the estimator of μ_1, say $\hat{\mu}_{10}$, is given by

$$\hat{\mu}_{10} = \left(n_T + \frac{n_R}{\delta^2}\right)^{-1} \left(n_T \bar{V} + \frac{n_R}{\delta} \bar{U}\right). \tag{2.12}$$

We shall implement the bootstrap calibration by generating parametric bootstrap samples $(\bar{V}^*, \bar{U}^*, S^{*^2})$ from the distributions

$$\bar{V}^* \sim N\left(\hat{\mu}_{10}, \frac{2S^2}{n_T}\right), \ \bar{U}^* \sim N\left(\hat{\mu}_{10}/\delta, \frac{2S^2}{n_R}\right), \text{ and } (n_T + 2n_R - 3)\frac{S^{*^2}}{S^2} \sim \chi^2_{n_T+2n_R-3}, \tag{2.13}$$

where S^2 is as specified in (2.1), and $\hat{\mu}_{10}$ is given in (2.12). Now let $(\bar{V}_i^*, \bar{U}_i^*, S_i^{*^2})$, i = 1, 2, ..., B, be a parametric bootstrap sample of size B generated from the distribution of $(\bar{V}^*, \bar{U}^*, S^{*^2})$ given in (2.13). Let

$$Z_i^* = \frac{|\frac{\bar{V}_i^*}{\bar{U}_i^*}| - \delta}{\frac{S_i^*}{\sqrt{n_T}} \sqrt{\frac{2}{\bar{U}_i^{*^2}} + \frac{4\bar{V}_i^{*^2}}{\bar{U}_i^{*^4}}}}, \tag{2.14}$$

i = 1, 2, ..., B. Our objective is to choose a γ that will make the type I error probability very close to α. For this, we can consider a grid of values of γ, and compute the proportion of times $Z_i^* < -z_\gamma$. Now pick a value of γ for which this proportion is equal to α. We used the *R* function *uniroot* to obtain γ. Let the resulting choice of γ be denoted by $\hat{\alpha}$ in order to emphasize that the significance level to be used is actually estimated from the data. The bootstrap calibrated test consists of rejecting H_0 in (1.2) when $Z < -z_{\hat{\alpha}}$, where Z is defined in (2.10).

2.2.2. *The gpq-based test*

As noted in Section 2.1, a gpq for θ is the quantity $\tilde{\theta}$ given in (2.4). In order to test (1.2) at significance level α, we compute the $100(1-\alpha)$th percentile of $\tilde{\theta}$, and reject H_0 if such a percentile is less than δ. Note that in order to compute such a percentile, we keep the observed data ($\bar{x}_T$, $\bar{x}_{R_1}$, $\bar{x}_{R_2}$ and s^2) as fixed, and treat the gpq as a random variable that is a function of the independent standard normal random variables Z_T, Z_{R_1} and Z_{R_2}, and the chisquare random variable $U^2 \sim \chi^2$ with df = $n_T + 2n_R - 3$. Several copies of the gpq can be generated by generating several values of (Z_T, Z_{R_1}, Z_{R_2}, U^2). The$100(1-\alpha)$th percentile of the gpq can be computed based on such generated values, which can then be used to carry out the test.

2.2.3. *Numerical Results*

In order to assess the performance of the delta method based test proposed by Kang and Chow [7] along with the improvement resulting from bootstrap calibration, and to compared with the gpq based test, we shall now report some simulation results. The sample sizes and parameter combinations used in the simulations are taken from Kang and Chow [7]. We used the sample sizes n_T = 30, 50, and 100, and $n_R = n_T/2$. Two values of δ were considered: δ = 1.1 and 1.2, and three values of σ were considered: σ = 1, 2, and 3. Table 1 gives the type I error probabilities of the test with rejection region (2.10), its bootstrap calibrated version, and the gpq based test. In the table, these are denoted by "Delta method", "Bootstrap calibration" and "GPQ Method", respectively. We have used a 5% nominal level, and the results in Table 1 are based on 10,000 simulations.

From the numerical results, it should be clear that the delta method based test exhibits poor performance when the sample size is small and/or when the common variance σ^2 is large. The bootstrap calibration improves the type I error performance of the test in terms of providing type I error probabilities close to the nominal level. However, the test is still not satisfactory. On the other hand, the gpq based test appears to provide satisfactory type I error probabilities in all scenarios considered for simulation. Clearly, the test to be recommended is the gpq based test.

Table 1. Type 1 error probabilities of the different tests in the equal variance case for a 5% significance level.

δ	μ_T	μ_{R1}	μ_{R2}	σ^2	n_T	Delta Method	Bootstrap Calibration	GPQ Method	μ_T	μ_{R1}	μ_{R2}	σ^2	n_T	Delta Method	Bootstrap Calibration	GPQ Method
					30	0.0631	0.0447	0.0441					30	0.0686	0.0535	0.0478
1.2	117	100	110	1	50	0.0564	0.0434	0.0524	110.2	106	100	1	50	0.0621	0.0499	0.0466
					100	0.0578	0.0477	0.0523					100	0.0581	0.0485	0.0543
					30	0.0647	0.0397	0.0476					30	0.0726	0.0519	0.0456
1.2	117	100	110	2	50	0.0557	0.0366	0.0499	110.2	106	100	2	50	0.0701	0.0523	0.0501
					100	0.0562	0.0441	0.0491					100	0.0583	0.0477	0.0511
					30	0.0694	0.0396	0.0453					30	0.0754	0.0492	0.0488
1.2	117	100	110	3	50	0.0606	0.0363	0.0451	110.2	106	100	3	50	0.0696	0.0509	0.0451
					100	0.0563	0.0404	0.0515					100	0.0630	0.0502	0.0479
					30	0.0590	0.0370	0.0461					30	0.0656	0.0490	0.0460
1.1	116	100	110	1	50	0.0592	0.0435	0.0498	109.6	106	100	1	50	0.0640	0.0527	0.0468
					100	0.0553	0.0455	0.0512					100	0.0605	0.0517	0.0470
					30	0.0637	0.0379	0.0457					30	0.0695	0.0478	0.0472
1.1	116	100	110	2	50	0.0614	0.0407	0.0492	109.6	106	100	2	50	0.0671	0.0531	0.0510
					100	0.0606	0.0435	0.0481					100	0.0615	0.0510	0.0457
					30	0.0646	0.0498	0.0468					30	0.0767	0.0510	0.0459
1.1	116	100	110	3	50	0.0659	0.0528	0.0476	109.6	106	100	3	50	0.0688	0.0497	0.0515
					100	0.0610	0.0530	0.0470					100	0.0644	0.0503	0.0497

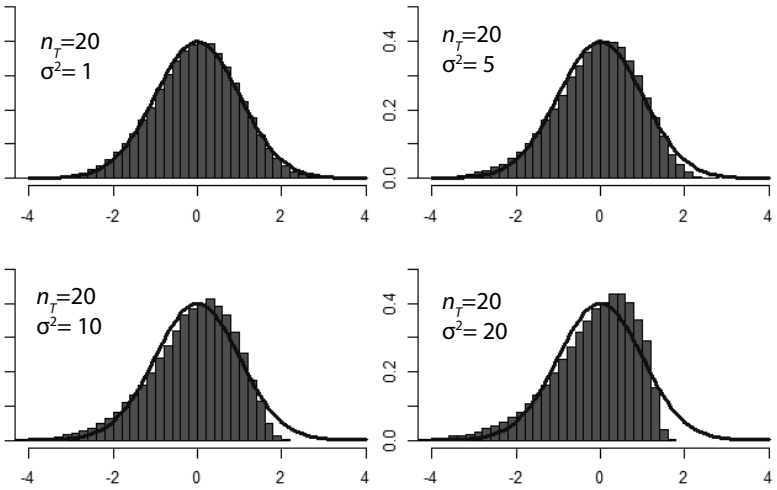

Fig. 1. Distribution of the test statistic under different variances and n_R =20

Table 2. Upper and Lower tail probabilities of the test statistic Z in (2.10) (based on 100,000 simulations)

σ^2	Probability	$\alpha = 0.05$			$\alpha = 0.1$		
		n_R			n_R		
		20	50	100	20	50	100
1	$P(Z < -z_\alpha)$	0.0592	0.0560	0.0545	0.1108	0.1085	0.1029
	$P(Z > z_\alpha)$	0.0382	0.0426	0.0453	0.0876	0.0911	0.0938
5	$P(Z < -z_\alpha)$	0.0721	0.0641	0.0596	0.1243	0.1143	0.1103
	$P(Z > z_\alpha)$	0.0227	0.0333	0.0379	0.0695	0.0813	0.0887
10	$P(Z < -z_\alpha)$	0.0785	0.0685	0.0641	0.1292	0.1190	0.1140
	$P(Z > z_\alpha)$	0.0095	0.0256	0.0331	0.0524	0.0743	0.0842

In order to have further insight into the poor performance of the delta method based test, we plotted a histogram of the test statistic based on 10,000 simulated values for $n_R = 20$, and $\sigma^2 = 1$, 5, 10 and 20. These appear in Figure 1, with the normal curve superimposed. It is clear that the normal approximation is poor as σ becomes large. In Table 2, we have given the tail probabilities of the distribution of the test statistic, below and above $-z_\alpha$ and z_α, where z_α is the upper α percentile of the standard normal distribution. The asymmetry of the distribution and the poor quality of the normal approximation should be clear from Table 2 when the sample size is small and/or σ^2 is large.

2.2.4. *Power comparison*

Figure 2 gives plots of the power curves of the three tests, plotted against the sample size, for σ^2 = 1. For this, the null value is chosen as $\theta = \delta = 1.2$ and the alternative used is $\theta = 0$. For each combination of the sample size and σ, the power was obtained using 5000×5000 simulations. We note that for small sample sizes, the delta method based test has a slightly larger power, and the gpq

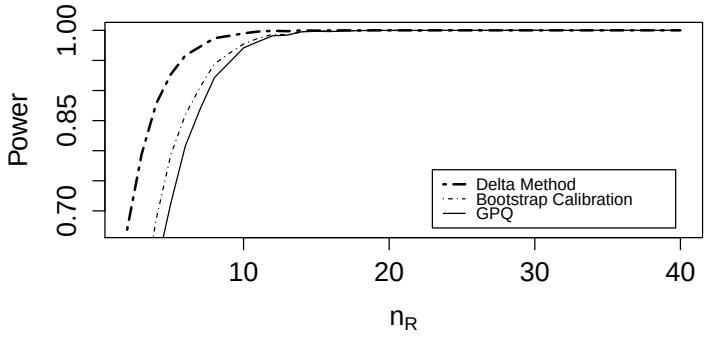

Fig. 2. Power functions plotted against the sample size for the alternative $\theta = 0$ when $\sigma_T^2 = \sigma_R^2 = \sigma^2 = 1$.

based test has somewhat lower power. This is to be expected since the type I error probabilities are slightly inflated for the delta method based test (as noted in Table 1). However, as the sample size gets large, the difference in power among the three tests disappears. More extensive plots are given in Pottackal [10], and the pattern noted in the different plots is the same as that in Figure 2.

2.3. *The case of unequal variances*

We now have the distributions specified in (2.2). A natural estimator of θ is once again given by $\hat{\theta}$ in (2.6). Defining

$$\mu_1 = \mu_T - (\mu_{R_1} + \mu_{R_2})/2,\ \mu_2 = \mu_{R_1} - \mu_{R_2},\ \sigma_1^2 = \frac{\sigma_T^2 + \sigma_R^2}{n_T} \text{ and } \sigma_2^2 = \frac{2\sigma_R^2}{n_R}, \tag{2.15}$$

$\hat{\theta}$ can also be expressed as

$$\hat{\theta} = \frac{\bar{V}}{\bar{U}},\ \text{ where } \bar{V} \sim N(\mu_1, \sigma_1^2) \text{ and } \bar{U} \sim N(\mu_2, \sigma_2^2), \tag{2.16}$$

where we have used the assumption $n_T = 2n_R$. Note that $\bar{V}$ and $\bar{U}$ are also independent. Once again, a straightforward application of the delta method gives

$$\sqrt{n_T}\left(\frac{\bar{V}}{\bar{U}} - \frac{\mu_1}{\mu_2}\right) \sim N\left(0,\ \frac{\sigma_1^2}{\mu_2^2} + \frac{\mu_1^2\sigma_2^2}{\mu_2^4}\right), \tag{2.17}$$

asymptotically. We reject H_0 when

$$\frac{|\frac{\bar{V}}{\bar{U}}| - \delta}{\sqrt{\frac{S_1^2}{\bar{U}^2} + \frac{\bar{V}^2 S_2^2}{\bar{U}^4}}} < -z_\alpha \tag{2.18}$$

where S_1^2 and S_2^2 are the variances given in (2.2). A bootstrap calibrated version can also be derived as done earlier for the case of equal variances. For this, we note that with σ_1^2 and σ_2^2 as defined in (2.15), respective estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are given by $\hat{\sigma}_1^2 = (S_T^2 + S_R^2)/n_T$ and $\hat{\sigma}_2^2 = 2S_R^2/n_R$, where

S_T^2 and S_R^2 have the distributions specified in (2.2). In view of (2.16), a derivation similar to (2.12) gives us

$$\hat{\mu}_{10} = \left(\frac{1}{\hat{\sigma}_1^2} + \frac{1}{\delta^2\hat{\sigma}_2^2}\right)^{-1}\left(\frac{\bar{V}}{\hat{\sigma}_1^2} + \frac{\bar{U}}{\delta\hat{\sigma}_2^2}\right).$$

In order to perform the bootstrap calibration, parametric bootstrap samples are generated from the distributions

$$\bar{V}^* \sim N\left(\hat{\mu}_{10}, \hat{\sigma}_1^2\right), \bar{U}^* \sim N\left(\hat{\mu}_{10}/\delta, \hat{\sigma}_2^2\right), (n_T-1)\frac{S_T^{*2}}{S_T^2} \sim \chi^2_{n_T-1}, (2n_R-2)\frac{S_R^{*2}}{S_R^2} \sim \chi^2_{2n_R-2}.$$

Furthermore, a gpq based test can also be derived. We recall that the derivation of the gpqs in the unequal variance scenario is explained in Section 2.1.

2.3.1. *Type I error and power*

Similar to Table 1 and Figure 2, Table 3 and Figure 3, respectively, give the type I error probabilities and power plots for the different tests in the unequal variance scenario for the sample sizes n_T = 30, 50, and 100, $n_R = n_T/2$, and null value δ = 1.2. Furthermore, the σ_T^2 and σ_R^2 were chosen to take the values 1, 2 and 3. The power plots have been obtained at the alternative value θ = 0, and is given only for the parameter choice σ_T^2 = 1 and σ_R^2 = 2 (see Pottackal [10] for further plots on the

Table 3. Type 1 error probabilities of the different tests in the unequal variance case for a 5% significance level.

δ	μ_T	μ_{R1}	μ_{R2}	σ_T^2	σ_R^2	n_T	Delta Method	Bootstrap Calibration	GPQ Method
1.2	117	100	110	1	1	30	0.0629	0.0380	0.0479
						50	0.0595	0.0409	0.0484
						100	0.0550	0.0429	0.0516
1.2	117	100	110	1	2	30	0.0687	0.0405	0.0509
						50	0.0634	0.0418	0.0468
						100	0.0579	0.0430	0.0498
1.2	117	100	110	1	3	30	0.0694	0.0393	0.0476
						50	0.0641	0.0412	0.0510
						100	0.0620	0.0442	0.0532
1.2	117	100	110	2	1	30	0.0618	0.0289	0.0463
						50	0.0579	0.0320	0.0491
						100	0.0584	0.0381	0.0497
1.2	117	100	110	2	2	30	0.0637	0.0332	0.0500
						50	0.0603	0.0353	0.0526
						100	0.0563	0.0391	0.0511
1.2	117	100	110	2	3	30	0.0682	0.0514	0.0491
						50	0.0622	0.0523	0.0508
						100	0.0604	0.0506	0.0474
1.2	117	100	110	3	1	30	0.0619	0.0301	0.0526
						50	0.0561	0.0266	0.0501
						100	0.0546	0.0266	0.0475
1.2	117	100	110	3	2	30	0.0657	0.0285	0.0487
						50	0.0610	0.0317	0.0500
						100	0.0572	0.0363	0.0479
1.2	117	100	110	3	3	30	0.0667	0.0302	0.0499
						50	0.0635	0.0352	0.0512
						100	0.0586	0.0391	0.0519

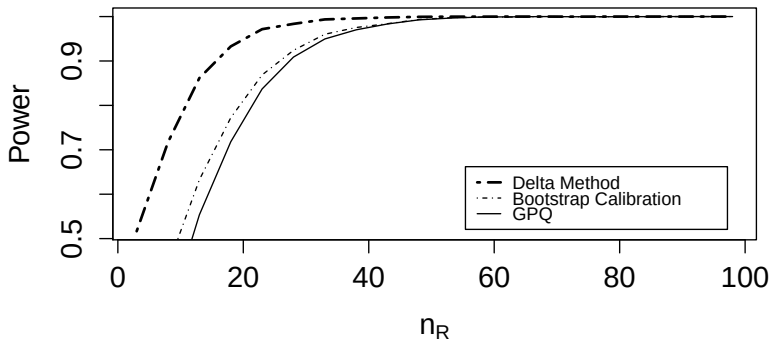

Fig. 3. Power functions plotted against the sample size for the alternative $\theta = 0$ when σ_T^2 and σ_R^2 are unequal, with σ_T^2 =1 and σ_R^2 = 2.

power). The conclusions from Table 3 and Figure 3 are similar to those for the equal variance case. In particular, the gpq based test emerges as the one that is most satisfactory.

2.3.2. *Sample size calculation*

In their paper, Kang and Chow [7] have provided table of sample sizes so that the delta method based test will provide 80% and 90% power. Table 4 gives the sample sizes n_T that will guarantee 90% power by the different tests in the unequal variance scenario.

Note that once n_T is determined, n_R is obtained as $2n_T$. From Table 4, we see that overall, the required sample size is slightly higher for the gpq based test. This is to be expected in view of the type I error performance of the tests.

Here we would like to point out that sample size determination in the context of equivalence trials is discussed in several articles; see [1] and [11]. Also, Kang and Kim [8] have addressed the sample size issue for biosimilar products; however, the set up considered by the authors is the usual bioequivalence scenario involving a single test drug and a single reference drug.

Table 4. Sample sizes n_T necessary to guarantee 90% power for the delta method based test, its bootstrap calibrated version and the gpq based test at the alternative value $\theta = 0$, for a 5% significance level.

σ_T^2	σ_R^2	Delta Method	Bootstrap Calibration	GPQ Method
1	1	5	8	8
2	2	9	12	13
3	3	12	18	21
1	2	19	27	29
1	3	23	37	39
1	4	29	49	51

3. Testing Average Biosimilarity Based on $|\mu_T - \mu_R| - |\mu_{R_1} - \mu_{R_2}|$

We note that the hypothesis in (1.2), formulated in terms of the parameter θ defined in (1.1), is what is given in Kang and Chow [7]. We have simply followed their formulation, and have improved their test using bootstrap calibration, suggested a test based on the gpq, and have also extended their results to the unequal variance scenario. It should however be noted that if the means μ_{R_1} and μ_{R_2} are very close, θ defined in (1.1) will have its denominator close to zero. This could present practical difficulties; for example, how can one choose the threshold δ in the hypotheses in (1.2), when μ_{R_1} and μ_{R_2} are unknown and their difference could be close to zero? It appears that an alternative formulation could mitigate this problem. Thus, instead of the ratio used to define θ in (1.1), we shall consider the parameter to be the difference defined in (1.3) and consider the hypotheses

$$H_0 : \theta_1 \geq \delta_1, \text{ vs } H_1 : \theta_1 < \delta_1, \tag{3.1}$$

for some specified threshold δ_1. Note that θ_1 also measures the magnitude of $|\mu_T - \mu_R|$ relative to that of $|\mu_{R_1} - \mu_{R_2}|$. We shall develop a test for the hypotheses in (3.1) using the gpq approach. Let $\tilde{\mu}_T$, $\tilde{\mu}_{R_1}$ and $\tilde{\mu}_{R_2}$, respectively denote gpqs for μ_T, μ_{R_1} and μ_{R_2}, as defined in Section 2.1. We recall that the definition of these quantities depend on whether the variances σ_T^2 and σ_R^2 are equal or unequal. Also define $\tilde{\mu}_R = (\tilde{\mu}_{R_1} + \tilde{\mu}_{R_2})/2$. A gpq for θ_1 defined in (1.3) is the quantity $\tilde{\theta}_1$ given by

$$\tilde{\theta}_1 = |\tilde{\mu}_T - \tilde{\mu}_R| - |\tilde{\mu}_{R_1} - \tilde{\mu}_{R_2}|, \tag{3.2}$$

and the hypotheses in (3.1) can be tested using the $100(1-\alpha)$th percentile of the resulting gpq. An estimate of this percentile can be easily obtained using Monte Carlo simulation, and we reject H_0 if the $100(1-\alpha)$th percentile so obtained is less than δ_1. Table 5 and Table 6 give the type I error probabilities of the resulting test in the equal variance and unequal variance scenarios. It should be clear that the gpq based test is quite accurate.

Table 5. Type 1 error probabilities of the gpq based test for testing the hypotheses in (3.1) at a 5% significance level ($\sigma_T^2 = \sigma_R^2$).

μ_T	μ_{R1}	μ_{R2}	σ_T^2	n		
				30	50	100
117	100	110	1	0.0522	0.0488	0.0516
			2	0.0465	0.0502	0.0500
			3	0.0499	0.0486	0.0463
116	100	110	1	0.0477	0.0479	0.0499
			2	0.0454	0.0512	0.0525
			3	0.0500	0.0479	0.0493
110.2	106	100	1	0.0463	0.0435	0.0519
			2	0.0465	0.0464	0.0522
			3	0.0499	0.0477	0.0526
109.6	106	100	1	0.0510	0.0475	0.0548
			2	0.0469	0.0491	0.0490
			3	0.0491	0.0482	0.0507

Table 6. Type 1 error probabilities of the gpq based test for testing the hypotheses in (3.1) at a 5% significance level (σ_T^2 and σ_R^2 are unequal and $\sigma_T^2 = 1$).

μ_T	μ_{R1}	μ_{R2}	σ_R^2	n		
				30	50	100
117	100	110	1	0.0459	0.0510	0.0475
			2	0.0534	0.0497	0.0492
			3	0.0524	0.0502	0.0461
116	100	110	1	0.0432	0.0516	0.0520
			2	0.0486	0.0496	0.0501
			3	0.0479	0.0483	0.0483
110.2	106	100	1	0.0485	0.0503	0.0471
			2	0.0484	0.0510	0.0504
			3	0.0495	0.0506	0.0502
109.6	106	100	1	0.0473	0.0529	0.0509
			2	0.0478	0.0491	0.0503
			3	0.0486	0.0527	0.0468

4. An Example

The example is taken from the European Medicines Agency document [5]. The problem addressed in the document is that of testing biosimilarity of Accofil, a biologic that is expected to be similar to the reference product Neupogen. The biotechnology drug Neupogen has been approved for use to boost white blood cell production in patients undergoing chemotherapy for certain cancers. Two versions, say R_1 and R_2, of Neupogen are used in the biosimilarity study: EU-approved Neupogen and US-licensed Neupogen, corresponding to two arms of the study, and the third arm corresponds to the biosimilar version Accofil (the test drug T). The data are actually generated using a three-period crossover design. For the purpose of illustrating our methodology, we shall ignore this aspect, and proceed with the assumption of a three-arm parallel design. The sample sizes are $n_T = 43$ for Accofil, and $n_R = 43$ for each reference arm. The observed values of the summary statistics (based on the AUC data) are

$$\bar{x}_T = 200720.00,\ \bar{x}_{R_1} = 192379.97,\ \bar{x}_{R_2} = 186404.48, s_T = 68,244.80,\ \text{ and }\ s_R = 60611.94.$$

The details given in the European Medicines Agency document [5] indicate that log-normality is reasonable. Since the summary statistics based on the log-transformed data are not available, we proceed assuming normality. The gpq based upper confidence limit for θ came out to be 15.92. Thus if we choose a value $\delta = 1.20$, the null hypothesis in (1.2) cannot be rejected. In other words, we cannot conclude biosimilarity.

5. Discussion

There is increasing interest in the development of biosimilars, and statistical criteria for establishing biosimilarity are still emerging. In the investigation of such criteria, two challenges present themselves: (i) the appropriate criterion/criteria that should be used to establish biosimilarity, and (ii) the development of accurate tests based on the chosen criteria. In this article, we have not advocated any specific statistical criterion; indeed, there is no consensus yet on the statistical criterion that should be used. However, we feel that average biosimilarity is a minimum requirement for a biosimilar

product. Average biosimilarity has already been investigated in the literature, and can be addressed in the context of two-arm and three-arm parallel designs. The latter set up has recently been taken up by Kang and Chow [7], and our work is also in the same scenario. The contribution in our work is three-fold: we have relaxed the equal variance assumption in their article, we have derived an accurate test using generalized pivotal quantities, and we have suggested an alternative formulation for assessing average biosimilarity. Our test procedures exhibit satisfactory performance in terms of type I error probabilities. Our overall conclusion is that average biosimilarity can be assessed accurately, even when the test and reference formulations exhibit different variabilities.

References

[1] Y.-W. Chang, Y. Tsong, X. Dong and Z. Zhao (2014). Sample size determination for a three-arm equivalence trial of normally distributed responses. Journal of Biopharmaceutical Statistics, **24**, 1190-1202.
[2] S.-C. Chow (2013). Biosimilars: Design and Analysis of Follow-on Biologics, Chapman & Hall/CRC Press, Boca Raton, Florida.
[3] S.-C. Chow and J. P. Liu (2008). Design and Analysis of Bioavailability and Bioequivalence Studies, Third Edition, Chapman & Hall/CRC Press, Boca Raton, Florida.
[4] B. Efron and R. Tibshirani (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC Press, Boca Raton, Florida.
[5] EMA (2014). CHMP Assessment Report: Accofil, Available at http://www.ema.europa.eu.
[6] FDA (2015). Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. The U.S. Food and Drug Administration, Silver Spring, MD.
[7] S. H. Kang and S.-C. Chow (2013). Statistical assessment of biosimilarity based on relative distance between follow-on biologics. Statistics in Medicine, **32**, 382-392.
[8] S. H. Kang and Y. Kim (2014). Sample size calculations for the development of biosimilar products. Journal of Biopharmaceutical Statistics, **24**, 1215-1224.
[9] K. Krishnamoorthy and T. Mathew (2009). Statistical Tolerance Regions: Theory, Applications and Computation, John Wiley & Sons, New York.
[10] Pottackal, G. J. (2015). Some Tests, Confidence Limits and Tolerance Limits for Assessing Biosimilarity. Doctoral dissertation submitted to the University of Maryland Baltimore County.
[11] A. Sun, X. Dong and Y. Tsong (2014). Sample size determination for equivalence assessment with multiple endpoints. Journal of Biopharmaceutical Statistics, **24**, 1203-1214.
[12] S. Weerahandi (1993). Generalized confidence intervals. Journal of the American Statistical Association, **88**, 899-905.
[13] S. Weerahandi (1995). Exact Statistical Methods for Data Analysis. New York: Springer.
[14] S. Weerahandi (2004). Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models, John Wiley & Sons, New York.