**Simultaneous Confidence Interval Methods for Analytical Similarity Assessment**

Jiayin Zheng1, Donglei Yin2, Mengdie Yuan3, Shein-Chung Chow3

1Biostatistics Program, Public Health Sciences, Fred Hutchinson Cancer Research Center,

1100 Fairview Ave N, Seattle, WA 98109, USA.

2Department of Applied Mathematics and Statistics, Stony Brook University,

100 Nicolls Road, Stony Brook, NY 11794, USA.

3Office of Biostatistics, Center for Drug Evaluation and Research,

U.S. Food and Drug Administration,

10903 New Hampshire Avenue, Silver Spring, MD 20993, USA.

**Abstract**

Analytical similarity assessment is the foundation of the development of biosimilar drug product, where the quality attributes to characterize the test product and the reference product needs to be shown statistically similar. When there were multiple references, e.g., a US-licensed reference product and a EU-approved reference, in addition to the similarity of the test product with each of the reference product, extra evidence for the similarity between the two reference products is also needed in the analytical similarity assessment. The method of pairwise comparisons has been widely used, but is recently criticized due to the lack of accuracy and reliability of each pairwise comparison since each comparison does not fully utilize all data collected from the three groups. In addition, since the equivalence criterion for analytical similarity is based on the variability of reference product, the pairwise method therefore uses different equivalence criteria in the three comparisons. To avoid these issues, we proposed an alternative method using simultaneous confidence approach based on the fiducial inference theory. Scenarios with and without the assumption of equal variance between the three products were discussed. For each scenario, three versions of simultaneous confidence approach were proposed based on the different assumptions of the population variance, and within each version, two types of simultaneous confidence interval were proposed. We then conducted extensive simulation studies to compare the performance of our proposed method and the pairwise method, and provided examples where the pairwise comparison approach failed but the simultaneous confidence approach passed to illustrate the concern of using pairwise method. The simulation result shows that the methods using the original version and integrated version of simultaneous confidence interval have significant larger power compared to the pairwise comparison method and meanwhile can well control the type I error rate. While the method using the least favorable version of simultaneous confidence interval demonstrates the smallest power among the four methods, thus is a conservative approach which is preferred for avoiding false positive conclusions.

**Keywords:** biosimilarity; multiple references; simultaneous confidence interval; fiducial inference

**1 Background**

When an innovative biological drug product is going off patent protection, biotechnology and/or pharmaceutical companies (sponsors) may seek regulatory approval for similar biological (biosimilar) products to the innovative product in European Union (through EMA) or the United States (through FDA). Thus, for assessment of biosimilarity between a proposed biosimilar product (test product) and an innovative biological product (reference product), there may be multiple references, e.g., a US-licensed reference product and a EU-approved reference product of the same product. When multiple references exist, the sponsors often obtain extensive analytical data intended not only to support a demonstration that the proposed biosimilar product and the US-licensed reference product are highly similar, but also to provide a justification of the relevance of the comparative data (e.g., pharmacokinetic and/or clinical data) generated using EU-approved reference to support a demonstration of biosimilarity of the proposed biosimilar product to the US-licensed reference product.

In practice, however, the following questions often encountered. Suppose there are two reference products: a US-licensed reference product and an EU-approved reference product. First, we may successfully demonstrate the proposed biosimilar product is highly similar to each of the two reference products, but fail to demonstrate that the two reference products are highly similar. Second, we are able to demonstrate that the proposed biosimilar product is highly similar to one of the two reference products but not the other. Third, it is an interesting question whether the two reference products should be combined (e.g., taking the average or adjust for their corresponding variability associated with the responses) for an overall biosimilarity assessment. To address the first two questions, the method of pairwise comparisons in conjunction with a head-to-head graphical comparison is often considered. For the third question, Kang and Chow (2013) proposed a three-arm study design for biosimilarity assessment under a various scenarios of criteria related to multiple references.

At the Oncologic Drugs Advisory Committee (ODAC) meeting on July 13th, 2017 for review of biosimilar products of Avastin and Herceptin, the method of pairwise comparisons has been criticized. When two reference products (e.g., a US-licensed reference and a EU-approved reference) were considered, the pairwise method includes three comparisons (i.e., a proposed biosimilar product versus a US-licensed reference product, the proposed biosimilar product versus an EU-approved reference product, and the US-licensed reference product versus the EU-approved reference product). The first criticism is related to lack of accuracy and reliability of each pairwise comparison since each comparison does not fully utilize all data collected from the three groups. In addition, since the equivalence criterion for analytical similarity is based on the variability of reference product, the pairwise method therefore uses different equivalence criteria in the three comparisons, thus may lead to inconsistent conclusions regarding the assessment of biosimilarity. Alternatively, the ODAC suggested the potential use of simultaneous confidence approach, which has the advantages of utilizing all data collected from the study and using consistent equivalence criterion.

In the next section, the method of pairwise comparisons for analytical similarity assessment with multiple references was briefly outlined. Next, in section 3, the proposed simultaneous confidence interval approach based on the fiducial inference theory was described, where scenarios with and without the assumption of equal variance between the three products were discussed. Besides, examples where the pairwise comparison approach failed but the simultaneous confidence approach passed were provided to illustrate the concern of using pairwise method. In section 4, extensive simulation studies were further conducted to compare the performance of our proposed method and the pairwise method. In section 5, Kang and Chow’s method for addressing the third question is discussed. Some concluding remarks are given in the last section.

**2 Method of Pairwise Comparisons**

**2.1 Equivalence Test for Tier 1 CQAs**

For CQAs in Tier 1, FDA recommends that an equivalent test can be performed to assess of analytical similarity. [FDA, 2017] As indicated by the FDA, a potential approach could be a similar approach to confidence interval method of bioequivalence testing for generic products under the raw data model. In other words, for a given CQA, we may test for equivalence by the following interval (null) hypothesis:

(1)

Where is the equivalence limit (or similarity margin), and and are the mean responses of the test (the proposed biosimilar) product and the reference product lots, respectively. Analytical equivalence (similarity) is concluded if the null hypothesis of nonequivalence (dissimilarity) is rejected. Under the above null hypothesis, analytical similarity would be accepted for a given CQA if the two-sided confidence interval of the mean difference is within

FDA further suggested that the equivalence acceptance criterion (EAC) as , where is the population standard deviation associated with the reference product. In practice, can be estimated based on test values of some randomly sampled lots from a pool of reference lots. The suggested EAC margin is considered as fixed margin conditioned on the observed test values from different reference lots. In equivalence test for CQAs from Tier 1, it is very challenging for the sponsors and/or biostatisticians when there are only a limited number of lots available (for both reference product and test product). Thus, it is suggested that the sponsors provide a plan on how the reference standard deviation, , would be estimated with satisfactory scientific/statistical justification.

For a given CQA in Tier 1, denote as the mean difference. Then null hypothesis (1) can be rewritten as:

(2)

Suppose there are nR reference lots and nT test lots for the equivalence test. Based on a two one-sided tests procedure, similarity is concluded if the null hypothesis of dissimilarity is rejected at the α level of significance, if

,

and

where is an estimator of Δ, zα is the lower α quantile of the standard normal distribution, and is an estimator of . The statistical method is based on the assumption that , where is the population standard deviation associated with the test product. For estimating , FDA recommends testing one sample from each reference lot for obtaining an estimator of . This approach is an unbiased estimate of . is the difference of the arithmetic means between the test samples and reference samples.

Note that since a two one-sided tests procedure is operationally equivalence to a () confidence interval approach in many cases, similarity is concluded if the () confidence interval falls within the limits of , ).

**2.2 Pairwise Comparisons with Multiple References**

Where there are multiple references, e.g., a US-licensed reference product and an EU-approved reference product of the same product, it is suggested pairwise comparisons be considered not only to (1) check whether the two reference products are highly similar, but also to (2) compare the proposed biosimilar with each of the two references.

Denote T, R1 and R2 as the proposed biosimilar (test) product, the first reference product (e.g., a US-licensed reference product), and the second reference product (e.g., an EU-approved reference product), respectively. The pairwise comparisons deal with the following three sets of interval hypotheses:

(3)

(4)

(5)

where the first two hypothesis use R1 as the reference and the third uses R2 as the reference. Each null hypothesis, i.e., (3)-(5) can be tested using the two one-sided tests procedure at the level of significance described in the previous section. As indicated earlier, since the two one-sided tests procedure is operationally equivalence to a () confidence interval approach in many cases, similarity is often concluded if the () confidence interval falls within the equivalence limit. Intuitively, pairwise comparisons sound reasonable. However, as indicated by the ODAC (Oncologic Drugs Advisory Committee) panel at the 2017 July 13th ODAC meeting, pairwise comparisons may not be justifiable due to the following deficiencies.

First, the equivalence limits may be different from one comparison to another. As it can be seen from hypotheses (3)-(5), hypotheses (3) and (4) use R1 as the reference product, which hypothesis uses R2 as the reference product. As a result, pairwise comparisons may be biased because the equivalence limits are data-driven which depend upon an estimated variability associated with the reference product. This may present critical issue in assessing biosimilarity especially when the test product is highly similar to each of the reference product but there is notable difference between the two reference products (i.e., the two reference product fail to pass the equivalence test) is observed.

The other criticism of pairwise comparisons is that each pairwise comparison does not fully utilize all data collected from the three treatment groups. That is, hypothesis (3) uses data obtained from both R1 and R2, hypothesis (4) is tested based on data from the test (T) product and the first reference product (R1), while hypothesis (5) considers data obtained from the test (T) product and the second reference (R2). This may present critical issue in assessing biosimilarity when there is evidence of heterogeneity in mean and/or variance among the three groups with limited number of lots (both test and/or reference lots) available.

As a result, the feasibility and/or validity of pairwise comparisons have been challenged.

**3 Simultaneous Confidence Approach**

As an alternative to the pairwise comparisons approach, the ODAC panel suggested the potential use of simultaneous confidence approach, which allows us to fully utilize all data from the study with single reference product (e.g., the US-licensed product). In this section, we will describe the simultaneous confidence interval method under a parallel-group design for analytical studies.

**3.1 Assumptions and Statistical Framework**

For illustration of the concept of simultaneous confidence interval and for simplicity, we will consider the case where there are one test product and two reference products, denoted by T, R1, and R2. Without loss of generality, let T, R1, and R2 be the test (proposed biosimilar) product, the US-licensed product, and the EU-approved product. We further assume that R1 is the primary reference product and R2 is the secondary reference product for regulatory submission.

For a given critical quality attribute (CQA), FDA recommends performing a single test on each lot. Let n1 be the samples from the n1 (primary) reference lots and let n2 be the samples from the n2 (secondary) reference lots. Test results from these samples are then used to obtain estimates of and , where and are the standard deviations associated with the primary reference product and secondary reference product, respectively. Furthermore, denote by the standard deviation associated with the test product. Now suppose there are , and lots for the test product, the primary reference product, and the secondary product, respectively. For a given test (primary reference, secondary reference) lot, assume that the test value follows a normal distribution with mean () and variance (). For equivalence test for CQAs in Tier 1, FDA’s recommended approach assumes that

and for

and for

and

and for

In other words,

,

,

and

, .

Thus,

and

lots are used for testing hypotheses (3)-(5) with estimates (based on the test values) of and . These estimates are then considered as the true values for obtaining the EAC margins.

Following the sampling plan of one sample from each reference lot as recommended by the FDA, the empirical variance estimators of and , denoted as and , respectively, follow the probability distributions below

(6)

(8.6

where and are Chi-square distributions with the degree of freedom and , respectively. For testing hypotheses with and obtained, denote and as the observations (test results) of the CQA in Tier 1 of the test arm, the primary reference arm and secondary reference arm, respectively.

To propose the simultaneous confidence interval methods under the framework described above, Zheng and Chow (2018) considered the scenarios of (i) under the assumption that and (ii) without the assumption that , which are briefly described below. Those proposed methods are all based on fiducial inference [Fisher, 1935; J Zheng et al. 2017] by calculating corresponding fiducial probabilities.

**3.2 Simultaneous Confidence Interval with the Assumption that**

Assume and samples for each arm are independent and identical distributed. Denote

,

and

.

We have

, and .

Follow similar idea of fiducial inference theory, the marginal fiducial distributions of the three location parameters can be obtained as follows:

Denote and as the probability density functions of the above three normal distributions, respectively. Since the three groups of samples,

and

,

are statistically independent between each other, the joint fiducial probability density function of can be express as . Now we define the first version of fiducial probability.

(7)

If the above , where is the pre-specified confidence level, the null hypothesis of (3) is rejected and analytical similarity between T and R1 is concluded.

As indicated earlier, two one-sided tests procedure is operationally equivalent to the confidence interval approach in many cases. Under (7), we propose the following two types of simultaneous confidence interval for namely type I restricted simultaneous confidence interval (RSCI I) and type II restricted simultaneous confidence interval (RSCI II), which are briefly outlined below

**Type I Restricted Simultaneous Confidence Interval (RSCI I)** For any , we first calculate the following fiducial probability based on

which is denoted as . When , is equal to . For any , we then find the minimal that satisfies

Denote the minimal by if it exists. Then the type I restricted simultaneous confidence interval (RSCI I) of can be obtained as , with the confidence level of q. If exists and , the analytical similarity between T and R1 is concluded. In other words, in this case, we have

.

**Type II restricted simultaneous confidence interval (RSCI II)** For any , the type II restricted simultaneous confidence interval (RSCI II) can be obtained similarly. We first calculate the follows fiducial probability based on

which is denoted as . When , is equal to . For any , find the minimal satisfying

Denote the minimal by if it exists. The RSCI II confidence interval of can be obtained as , with the confidence level of q. If exists and , the analytical similarity between T and R1 is concluded. In this case, we have

.

Note that in practice, the true value of is often unknown. In this case, we can simply replace by its estimate in all of expressions above and obtained estimates for the fiducial probability in (8.7), i.e., and the two restricted simultaneous confidence intervals. (i.e., , ). In practice, if is a good estimate of , it is expected that and would perform similarly as compared with the RSCI assuming that is known.

It can be easily verified that and . Thus, RSCI II confidence interval approach is more conservative than RSCI I confidence interval approach. In other words, RSCI I confidence interval tends to, more favorably, conclude the rejection of all of the hypotheses as compared to that of RSCI II confidence interval.

**Modified RSCI I and RSCI II Confidence Intervals**

As discussed in the previous section, () is considered as know (its estimate is fixed as the true value). However, in real world, is often unknown and there exists variability associated with the estimate of (i.e., ). To take this variability into consideration, Zheng and Chow (2018) also proposed two modified simultaneous confidence intervals based on , and . One is referred to as the integrated version and the other is known as the least favorable version. Both modified simultaneous confidence intervals are derived based on the fiducial distribution of given in (6). As it can be seen from (6), the fiducial distribution of can be expressed as

,

where is considered as fixed and is Chi-square distribution with degree of freedom . Denote the probability density function of this fiducial distribution as e1.

***The integrated version*** The integrated fiducial probability (IFP) can be expressed as

(8)

Similarly, replace

and

simply by their integrated versions

and

in the expressions above. Then with the same derivation, we have the type I integrated restricted simultaneous confidence interval (IRSCI I) for and the type II integrated restricted simultaneous confidence interval (IRSCI II) for .

***The least favorable version*** it would be more conservative when the used value for is smaller, i.e., it’s hard to reject all three hypotheses with smaller value of . Thus, we suggest another version using the lower fiducial confidence bound to estimate the least favorable values of , i.e.,

where is the quantile of Chi-square distribution with degree of freedom . This leads to the least favorable fiducial probability (LFFP) , the type I least favorable restricted simultaneous confidence interval (LFRSCI I) for and the type II least favorable restricted simultaneous confidence interval (LFRSCI II) for .

**3.3 Simultaneous Confidence Interval without the Assumption of**

Now we do not assume but still assume that samples for each arm are independent and identical distributed. Two sets of methods are proposed: one uses only one reference (i.e., R1); the other uses two references (i.e., R1 and R2).

**3.3.1 The Case of One Reference**

Denote

and

.

We have

,

and

,

where is the t distribution with degree of freedom n-1. Follow similar idea of fiducial inference theory, the marginal fiducial distributions of the two location parameters can be obtained as follows:

Denote and as the probability density functions of the above two fiducial distributions, respectively. Since the three groups of samples,

and

,

are statistically independent between each other, the joint fiducial probability density function of can be express as . Now we define the second version of fiducial probability.

(9)

If , where is the pre-specified confidence level, all hypotheses in (3) are rejected and analytical similarity between T and R1 is concluded.

Based on (9), the following two types of simultaneous confidence interval of namely type III restricted simultaneous confidence interval (RSCI III) and type IV restricted simultaneous confidence interval (RSCI IV) can be similarly derived.

**Type III Restricted Simultaneous Confidence Interval (RSCI III)** Similarly, for any , we calculate the following fiducial probability based on

which is denoted by . Note that when , is equal to . For any , find the minimal that satisfies

Denote the minimal by if it exists. Then the type III restricted simultaneous confidence interval (RSCI III) of denoted by with the confidence level of q can be obtained. If exists and , the analytical similarity between T and R1 is concluded. In this case, we have

.

**Type IV Restricted Simultaneous Confidence Interval (RSCI IV)** To obtain a type IV restricted simultaneous confidence interval (RSCI IV), similarly, for any , calculate the fiducial probability based on

which is denoted as . When , is equal to . For any , then find the minimal satisfying

Denote the minimal by if it exists. Then we RSCI IV confidence interval of denoted by with the confidence level of q can be obtained. Thus, if exists and , the analytical similarity between T and R1 is concluded. In other words, we have

.

Similarly, we can replace by its estimate in all expressions and obtain estimated versions of the fiducial probability and the two restricted simultaneous confidence intervals, which are denoted by , and , respectively. Note that if is a good estimate of , it is expected that and would perform similarly as compared with the RSCI assuming that is known.

It can be easily verified that and . Thus, RSCI IV confidence interval is considered more conservative than RSCI III confidence interval.

**Modified RSCI III and RSCI IV Confidence Intervals**

To take the variability associated with the estimate of into consideration, two modified versions for , , and can be similarly derived. One is the integrated version and the other is the least favorable version. Both are based on the fiducial distribution of in (4). The fiducial distribution of can be expressed as , where is considered as fixed and is Chi-square distribution with degree of freedom . Denote the probability density function of this fiducial distribution as e1.

***The integrated version*** The integrated fiducial probability (IFP) can be expressed as

*(8)*

Similarly, simply replace

and

by their integrated versions

and

in the expressions above. Then with the same derivation, we have the type III integrated restricted simultaneous confidence interval (IRSCI III) for and the type IV integrated restricted simultaneous confidence interval (IRSCI IV) for .

***The least favorable version*** it would be more conservative when the used value for is smaller, i.e., it’s hard to reject all three hypotheses with smaller value of . Thus, we suggest another version using the lower fiducial confidence bound to estimate the least favorable value of , i.e.,

where is the quantile of Chi-square distribution with degree of freedom . This leads to the least favorable fiducial probability (LFFP) , the type III least favorable restricted simultaneous confidence interval (LFRSCI III) for and the type IV least favorable restricted simultaneous confidence interval (LFRSCI IV) for .

**3.3.2 The Case of Two References**

The above-proposed methods all use single variance reference for EAC. For one of the three hypotheses, , it may also be reasonable to use for EAC. To accommodate it, we propose another version of fiducial probabilities and the corresponding simultaneous confidence intervals with the fiducial density function .

*(9)*

If the above , where is the pre-specified confidence level, all hypotheses in (3) are rejected and analytical similarity between T and R1 is concluded.

In addition, we provide two types of simultaneous confidence interval of as follows.

***Type V restricted simultaneous confidence interval (RSCI V)*** For any , calculate the follows fiducial probability based on

denoted as . When , is equal to . For any , look for the minimal satisfying

Denote the minimal by if it exists. Then we get the type V restricted simultaneous confidence interval of as , with the confidence level of q. If exists and , the analytical similarity between T and R1 is concluded. In other words, in this case, we have

.

***Type VI restricted simultaneous confidence interval (RSCI VI)*** To obtain a type VI restricted simultaneous confidence interval (RSCI VI), similarly, for any , calculate the fiducial probability based on

which is denoted as . When , is equal to . For any , then find the minimal satisfying

Denote the minimal by if it exists. Then we RSCI VI confidence interval of denoted by with the confidence level of q can be obtained. Thus, if exists and , the analytical similarity between T and R1 is concluded. In other words, we have

.

Similarly, we can replace and by its estimate and in all expressions and obtain estimated versions of the fiducial probability and the two restricted simultaneous confidence intervals, which are denoted by , and , respectively. Note that if and are good estimates of and , it is expected that and would perform similarly as compared with the RSCI assuming that are known.

It can be easily verified that and . Thus, RSCI VI confidence interval is considered more conservative than RSCI V confidence interval.

**Modified RSCI V and RSCI VI Confidence Intervals**

To take this variability into consideration, two modified versions for above , and are also provided. One is the integrated version and the other is the least favorable version. Both are based on the fiducial distribution of and in (4). The fiducial distributions of and can be expressed as

and

,

respectively, where and are considered as fixed and is Chi-square distribution with degree of freedom . Denote the probability density function of the two fiducial distributions as e1 and e2.

***The integrated version*** The integrated fiducial probability (IFP) can be expressed as

*(10)*

Similarly, simply replace

and

with their integrated versions

and

in the expressions above. Then with the same derivation, we have the type V integrated restricted simultaneous confidence interval (IRSCI V) for and the type VI integrated restricted simultaneous confidence interval (IRSCI VI) for .

***The least favorable version*** it would be more conservative when the used values for and/or are smaller, i.e., it’s hard to reject all three hypotheses with smaller values of and/or . Thus, we suggest another version using the lower fiducial confidence bounds to estimate the least favorable values of and/or , i.e.,

,

where is the quantile of Chi-square distribution with degree of freedom . This leads to the least favorable fiducial probability (LFFP)

,

the type V least favorable restricted simultaneous confidence interval (LFRSCI V) for and the type VI least favorable restricted simultaneous confidence interval (LFRSCI VI) for .

For all the estimators proposed above, an accurate estimation can be obtained by numerical integration and solving a one-dimensional nonlinear equation.

**3.4 Illustration with examples**

Here we provide numeric examples to illustrate the concern of using pairwise method.

**Example 3.4.1 Pairwise method concludes false negative**

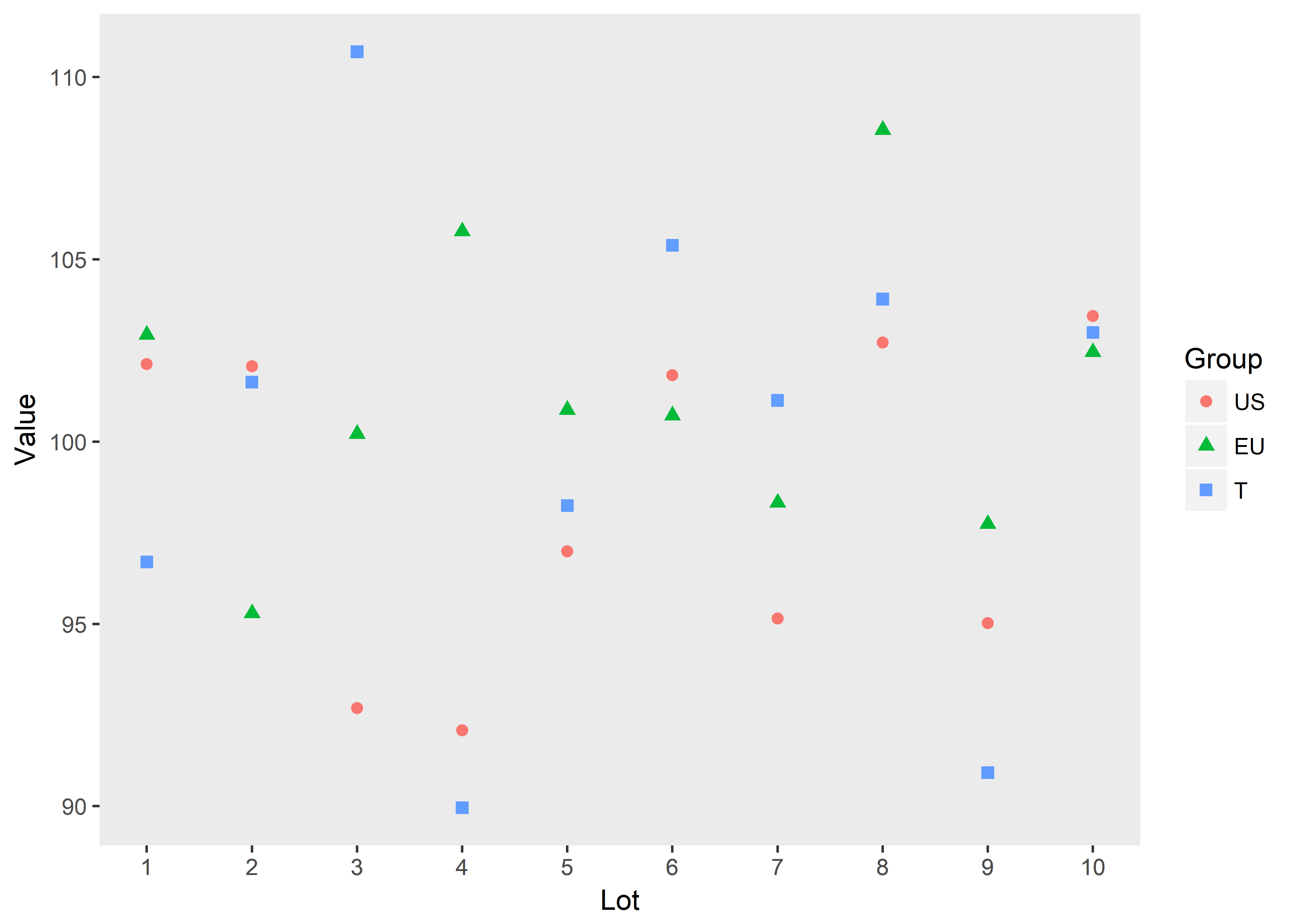
Suppose we have two reference products US reference and EU reference, denoted by US and EU, and one test product, denoted by T. Assume US, EU, and T follow normal distributions and share equal variance. The true means of the three products were set to be 99, 101, 100, and the true standard deviation was 6. Three groups of samples with equal size 10 were randomly generated from US, EU, and T population, respectively. Another two groups of samples with size 10 were randomly taken from the US and EU population to obtain the “true” standard deviations. The type I error allowed was set to be 0.1. Three pairwise comparisons, US versus EU, US versus T, EU versus T, were analyzed using the FDA recommended approach, with US, US, and EU as the references, respectively. The data were displayed in Table 1 and corresponding scatter plot was showed in Figure 1.

**Table 1. Random samples generated from the three population.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **Lot** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **US** | 102.13 | 102.07 | 92.69 | 92.09 | 96.99 | 101.83 | 95.15 | 102.72 | 95.02 | 103.45 |
| **EU** | 102.93 | 95.29 | 100.21 | 105.77 | 100.87 | 100.72 | 98.33 | 108.55 | 97.74 | 102.46 |
| **T** | 96.70 | 101.63 | 110.70 | 89.96 | 98.25 | 105.39 | 101.13 | 103.91 | 90.92 | 102.99 |
| **US (ref)1** | 104.38 | 99.39 | 102.71 | 103.81 | 95.35 | 102.41 | 97.56 | 101.56 | 95.10 | 99.96 |
| **EU (ref)1** | 101.39 | 104.90 | 98.09 | 98.32 | 101.82 | 107.23 | 83.62 | 100.30 | 106.98 | 98.52 |

1, samples randomly taken from the US and EU population to obtain the “true” standard deviations

**Figure 1. Scatterplots of the random samples generated for each group.**



**Table 2. The results of pairwise comparison approach vs. Simultaneous confidence interval approach.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pairwise comparison approach** | | | | | **Simultaneous confidence interval approach** | | | | |
| **Comparison** | **Mean difference** | **90% CI** | **EAC margin1** | **Equivalence Test** | **Method** | **Fiducial probability** | **Type 1 90% CI** | **Type 2 90% CI** | **Simultaneous**  **similarity** |
| EU vs. US | 2.87 | (0.42,5.33) | 5.01 | Fail | Original | 0.92 | (-4.51, 4.51) | (-4.79, 4.79) | Pass |
| T vs. US | 1.74 | (-0.72,4.20) | 5.01 | Pass | Integrated | 0.92 | (-4.51, 4.51) | (-4.84, 4.84) | Pass |
| T vs. EU | -1.13 | (-6.08,3.82) | 10.09 | Pass | Least Favorable | 0.79 | NA | (-4.15, 4.15) | Fail |

1, Similarity margin = 1.5\*sigma(Ref)

Under this setting, an effective test should be able to reject the null hypothesis and conclude the similarity. However, from Table 2, the pairwise comparison approach failed to reject one of the null hypotheses that the two reference drugs are not similar enough (EU vs. US, 90% CI: 0.42-5.33, exceeds the EAC margin=5.01). While two out of the three simultaneous confidence interval methods had fiducial probabilities calculated higher than 0.9 (0.92 for both original version and integrated version), and the corresponding two versions of confidence intervals lie within the simultaneous margin, thus could successfully reject all three hypotheses in (3), (4) and (5), i.e., conclude similarity among US, EU, and T. However, the least favorable version failed to conclude the similarity (fiducial probability=0.79). This example illustrates the case that pairwise method fail to conclude similarity when true similarity holds (i.e., false negative towards the hypothesis tests), and compared to this, the new proposed simultaneous interval approach was able to reject the null hypothesis and thus more powerful.

**Example 3.4.2 Pairwise method concludes false positive**

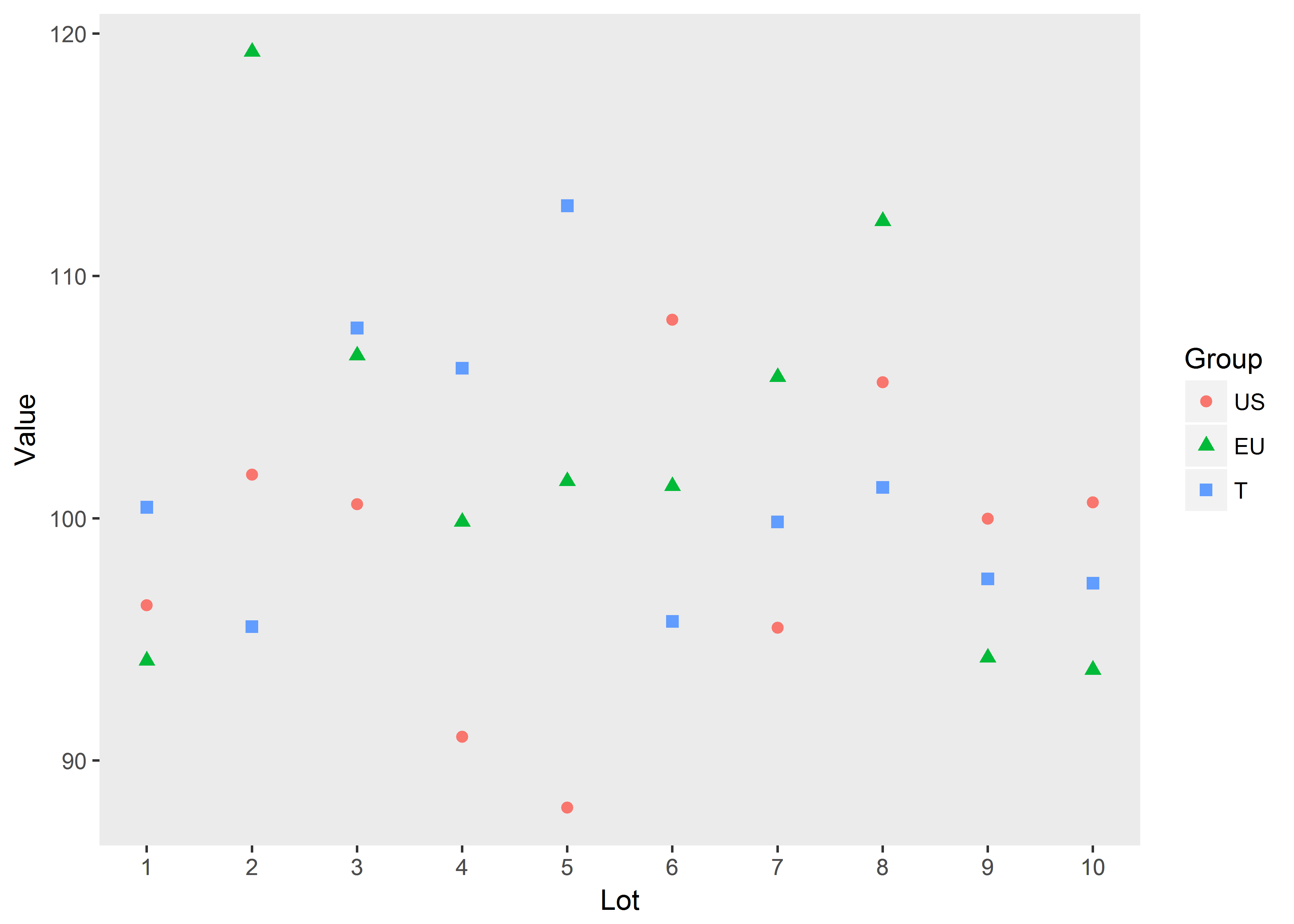
Suppose the true means of the three products was set to be 95, 105, 100, and the true standard deviation was 6. Similarly, three groups of samples with equal size 10 were randomly generated from US, EU, and T population, respectively. Another two groups of samples with size 10 were randomly taken from the US and EU population to obtain the “true” standard deviations. The type I error allowed was set to be 0.1. Three pairwise comparisons, US versus EU, US versus T, EU versus T, were analyzed using the FDA recommended approach, with US, US, and EU as the references, respectively. The data were displayed in Table 3 and corresponding scatter plot was showed in Figure 2.

**Table 3. Random samples generated from the three population.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **Lot** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **US** | 96.41 | 101.81 | 100.58 | 90.98 | 88.06 | 108.19 | 95.49 | 105.62 | 99.98 | 100.66 |
| **EU** | 94.13 | 119.26 | 106.72 | 99.86 | 101.54 | 101.33 | 105.83 | 112.27 | 94.25 | 93.75 |
| **T** | 100.46 | 95.53 | 107.85 | 106.19 | 112.90 | 95.75 | 99.85 | 101.27 | 97.50 | 97.32 |
| **US (ref)1** | 89.81 | 105.22 | 95.79 | 93.76 | 95.01 | 96.70 | 104.06 | 98.06 | 90.52 | 88.88 |
| **EU (ref)1** | 94.19 | 100.15 | 104.23 | 116.74 | 103.17 | 109.69 | 106.76 | 118.46 | 104.73 | 106.25 |

1, samples randomly taken from the US and EU population to obtain the “true” standard deviations

**Figure 2. Scatterplots of the random samples generated for each group.**



**Table 4. The results of pairwise comparison approach vs. Simultaneous confidence interval approach.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pairwise comparison approach** | | | | | **Simultaneous confidence interval approach** | | | | |
| **Comparison** | **Mean difference** | **90% CI** | **EAC margin1** | **Equivalence Test** | **Method** | **Fiducial probability** | **Type 1 90% CI** | **Type 2 90% CI** | **Simultaneous**  **similarity** |
| EU vs. US | 4.12 | (0.02,8.21) | 8.36 | Pass | Original | 0.94 | (-6.48, 6.48) | (-7.63, 7.63) | Pass |
| T vs. US | 2.68 | (-1.41,6.78) | 8.36 | Pass | Integrated | 0.95 | (-6.38, 6.38) | (-7.72, 7.72) | Pass |
| T vs. EU | -1.43 | (-6.74,3.88) | 10.83 | Pass | Least Favorable | 0.84 | NA | (-6.60, 6.60) | Fail |

1, Similarity margin = 1.5\*sigma(Ref)

Under this setting, an effective test should be able to accept the null hypothesis and not conclude the similarity. However, Table 4 indicates that follow the pairwise comparison approach, the data passes all the three equivalence tests and incorrectly concluded the similarity between three drug products (all the three 90% CIs lie within the EAC margins). Considering the simultaneous confidence interval approach, although the original and integrated versions of simultaneous confidence interval approaches also incorrectly concluded the similarity (had fiducial probabilities calculated higher than 0.9: 0.94 for original version and 0.95 for integrated version, and the corresponding two versions of confidence intervals lie within the simultaneous margin), the least favorable version successfully detected the difference and did not conclude the similarity (fiducial probability=0.79). This example illustrates the case that pairwise method incorrectly conclude similarity when significant difference between the three groups truly exists (i.e., false positive towards the hypothesis tests), and compared to this, the new proposed least favorable version of simultaneous interval approach was more conservative and avoided the type I error in this case. Further discussion of the new methods’ performance under different parameter settings can be found in the simulation studies of the following section.

**4. Simulation Studies**

Simulation studies were used to assess the performance of the proposed methods. For parameter specification, assume the three drugs, R1, R2 and T follow normal distributions with equal variance. the observations are statistically independent with each other. 10 lots were sampled for each drug, as well as for estimating the variance of the reference drug. The required type I error was set to be 0.1. A wide range of standard deviations (2/1.5, 2, 4, 6, 8, 10, 12, 14) were considered to represent different signal-noise ratios, of which 2/1.5 represents the margin between and . Margin determination was referred to the suggestion on page 10 of the 2017 FDA guidance [FDA, 2017]. For each standard deviation, 1000 repetitions were simulated. We discuss two scenarios where the means difference between the two reference drugs R1, R2 are closer than to the test drug T, as well as the other way around.

**Case 1:**

The expectations of the three drugs R1, R2 and T were set to be 99, 100 and 101, respectively. For summarization of the simulation results, we calculated the rate of rejecting all three hypotheses for each method (i.e., power, if above the margin or type I error, if on the margin), and the coverage rate of each simultaneous CI. Besides, the rate of rejecting all three hypotheses by the pairwise comparison approach was also calculated for comparison with the proposed methods. The summary results are shown in Table 5. We see all methods control the type I error no larger than the nominal level of 0.1. Two proposed simultaneous confidence interval approach (original and integrated version), performed significant better than the pairwise comparison method. For example, while standard deviation equals 8, the first two simultaneous confidence interval approach had power 0.74 and 0.72, respectively, much higher than the pairwise comparison method, which had overall power only 0.65, and the corresponding simultaneous confidence intervals, (-) and (-) had 100% coverage rate of the true means.

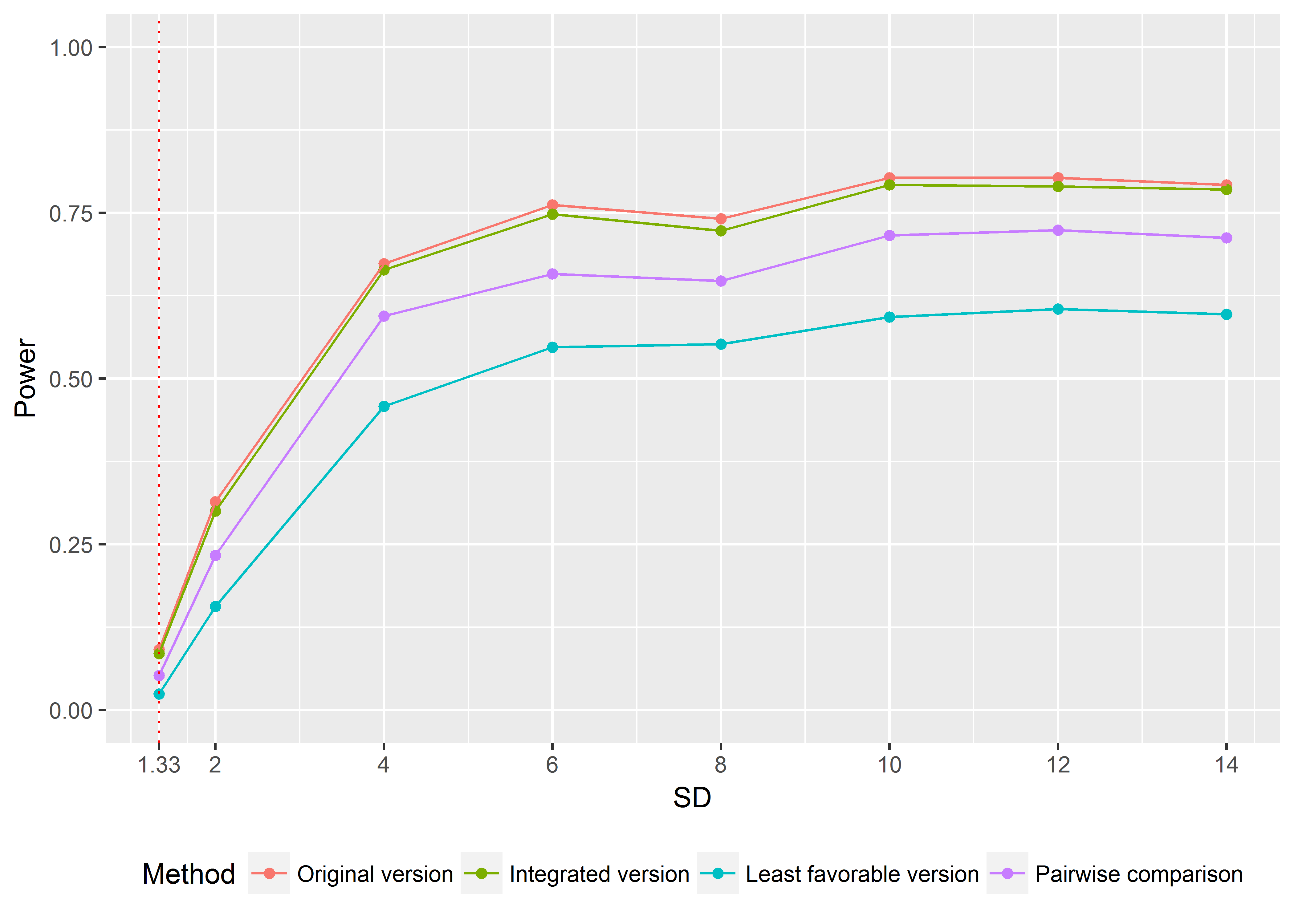
The least favorable version, compared to the rests, had the lowest power.

**Table 5. Simulation statistics of different methods ().**



Figure 3 further illustrates the comparison of power between pairwise comparison and the three simultaneous confidence interval methods. The purple line represents the pairwise comparison, and the rest three lines represents the three versions of simultaneous confidence interval approach (original, integrated and least favorable version). The x-axis represents the sample standard deviation which takes values from 1.33 to 14, where the red-dotted line (SD=1.33) is the rejection margin under EAC. The y-axis represents the empirical power based on 1000 replications. It indicates that the original (orange line) and integrated (green) version of simultaneous confidence interval approach have significantly larger power than the pairwise comparison (purple), and the superiority maintains as the standard deviance varies. While the least favorable version (blue) performs the worst among the four methods.

**Figure 3. Simulation performance of different methods ().**



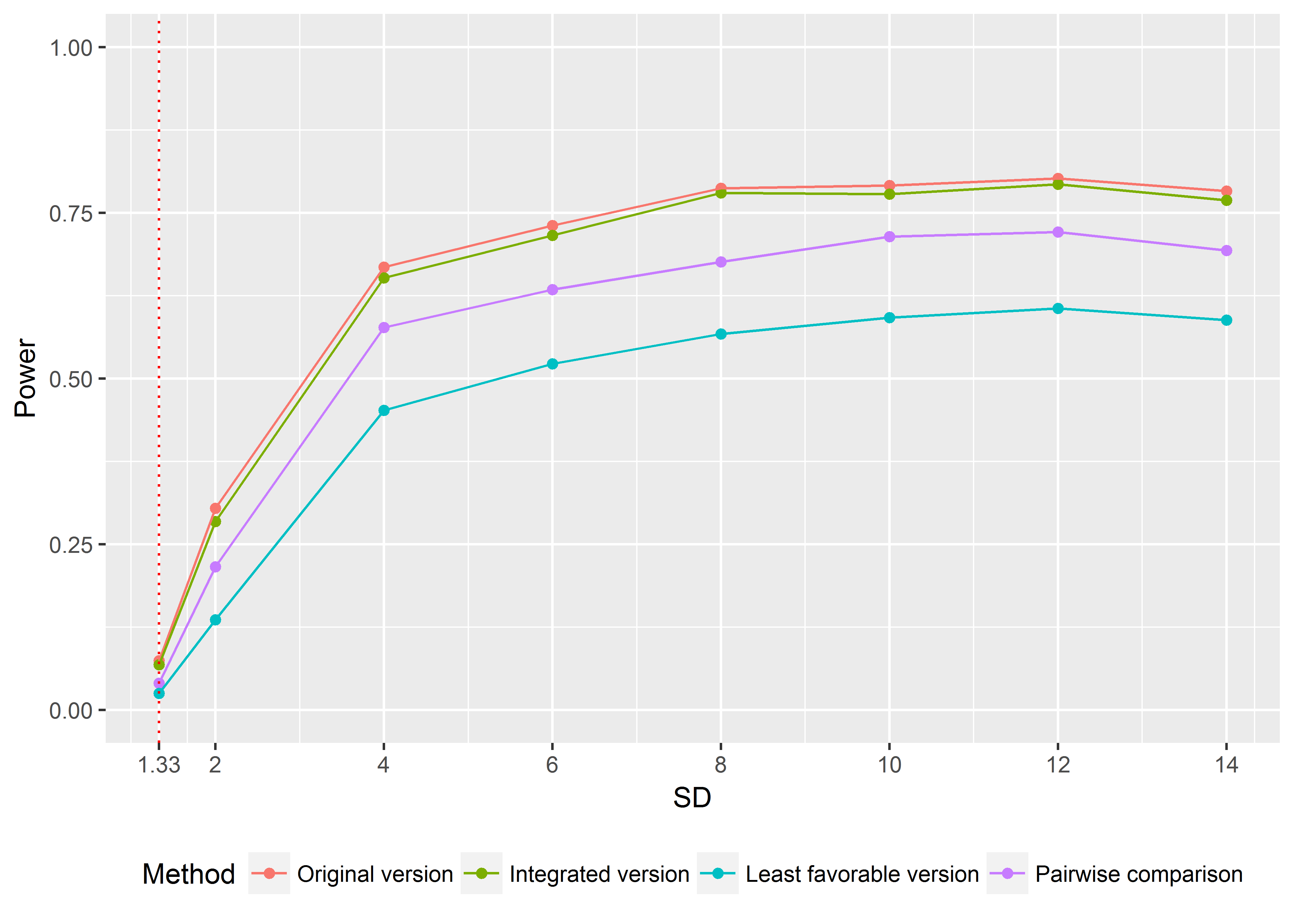
**Case 2:**

The expectations of the three drugs R1, R2 and T were set to be 99, 101 and 100, respectively. Similarly, From Table 6, we see all methods control the type I error no larger than the nominal level of 0.1. Two proposed simultaneous confidence interval approach (original and integrated version), had significant larger power than the pairwise comparison method. The least favorable version, compared to the rests, had the lowest power. Figure 4 further illustrates the comparison of power between pairwise comparison and the three simultaneous confidence interval methods under this setting.

**Table 6. Simulation statistics of different methods ().**



**Figure 4. Simulation performance of different methods ().**

**

**5. Concluding Remarks**

Current pairwise comparison method has been criticized due to the lack of accuracy and reliability of each pairwise comparison since each comparison does not fully utilize all data collected from the three groups, and the inconsistent use of different equivalence criteria in the three comparisons. To avoid these issues, we proposed an alternative method using simultaneous confidence approach based on the fiducial inference theory. We then provided examples where the pairwise comparison approach failed but the simultaneous confidence approach passed to illustrate the concern of using pairwise method, and conducted extensive simulation studies to further compare the performance of our proposed method and the pairwise method. The simulation result shows that the methods using the original version and integrated version of simultaneous confidence interval have significant larger power compared to the pairwise comparison method and meanwhile can well control the type I error rate. While the method using the least favorable version of simultaneous confidence interval demonstrates the smallest power among the four methods, thus is a conservative approach which is preferred for avoiding false positive conclusions.

**References**

FDA (2017). Guidance for Industry – Statistical Approaches to Evaluate Analytical Similarity. Food and Drug Administration, Silver Spring, Maryland, September, 2017.

Fisher RA. The fiducial argument in statistical inference. Annals of Human Genetics 1935; 6(4):391–398.

Jiayin Zheng, Shein‐Chung Chow, Mengdie Yuan. On Assessing Bioequivalence and Interchangeability between Generics Based on Indirect Comparisons. Statistics in Medicine, 2017, 36(19): 2978-­‐2993. DOI: 10.1002/sim.7326