A Novel Bootstrapping Test for Analytical Biosimilarity

Thomas Zahel1

Abstract

Showing analytical similarity is key to license biosimilar products with reduced or circumvented clinical effort. Statistical procedures to assess analytical similarity of quality attributes at drug product level have been highly debated by academia, industry, and regulatory agencies. In the past, a tiered approach was recommended by regulators, consisting of equivalence tests and quality range tests. However, this approach has recently been withdrawn by FDA. New guidelines of FDA and EMA favour the usage of quality range tests. Moreover, it has recently been shown that simple range tests, such as the 3SD test, are flawed, since they do not control the agency risk of falsely declaring a non-biosimilar product as being biosimilar (Type I error). This has also been highlighted by regulators recently. In this contribution, we developed a novel bootstrapping test for assessing analytical similarity that overcomes current flaws of equivalence and range tests. The developed test shows the desired properties, that is, (i) similarity conditions can be easily defined, (ii) differences of mean and the variance between the biosimilar and the innovator can be studied simultaneously, and (iii) the Type I error of the test can be controlled at a low level, e.g. 5%, evenly along the entire similarity condition. Moreover, the test shows up to 10% higher mean power values in the similarity region compared to existing range tests that aim to control the Type I error. Hence, this test is superior to existing quality range tests and is perceived compliant with current regulatory requirements.

Keywords analytical similarity · biosimilar · biosimilarity · bootstrapping · equivalence tests · range test

Introduction

Biosimilars are reaching increased clinical, regulatory, and commercial importance since innovator products are going off patent [1, 2]. During the development of biosimilar products (test product TP), similarity to the reference product (RP) needs to be shown. This can be done at multiple levels such as product quality attributes, pharmacokinetic (PK), animal, or clinical studies [3, 4]. Shown similarity at the quality attribute level at the end of the manufacturing process or PK level may serve as additional information or even surrogates to clinical studies expecting no meaningful differences in efficacy and safety of the product [5].

Regulatory Background

Recent regulatory efforts resulted in a series of guidance documents and reflection papers about statistical aspects of analytical biosimilarity assessment [6, 7]. The American Food and Drug Administration (FDA) requirements are less specific than European Medicines Agency (EMA) in their latest reflection paper [6]. In general, FDA states that “The objective of the comparative analytical assessment is to verify that each attribute, as observed in the proposed biosimilar and the reference product, has a similar population mean and similar population standard deviation.” More specifically, they propose to conduct a quality range approach, i.e. checking if a fraction of TP batches are within \( \bar{x}_{RP} \pm k \times s_{RP} \), where \( \bar{x}_{RP} \) represents the sample mean and \( s_{RP} \) the sample standard deviation of the reference product. The factor \( k \) can be adjusted depending on the criticality of the quality attribute; i.e. for more critical attributes a lower value of \( k \) might be chosen. Additionally the sponsor may use equivalence tests. For low-risk quality attributes, graphical comparison may be applied. It is important to note that the equivalence tests have been previously employed for the
most important quality attributes [8]. However, these equivalence tests are no longer highly recommended, neither by FDA nor by EMA.

Recent regulatory efforts of EMA resulted in a detailed reflection paper that describes the establishment of analytical biosimilarity. One important message from this reflection paper is that the process needs to be understood as a distribution from which we sample individual lots. Any claim on similarity/equivalence needs to be understood as a claim on the underlying distributions not on the actual samples. This is the basic concept of statistical inference. Additionally, inferential statistics require that the two distributions that are compared are representative for the true TP and RP manufacturing process, respectively. This is also covered in EMA and FDA guidelines; e.g. FDA states that TP lots should be “representative of the intended commercial manufacturing process” [6]. Therefore, we assume for the presented methods in this contribution that samples are representative and the following workflow can be extracted from [7]:

1. Define general aim

   • (non-inferiority or equivalence). This depends on the nature of the critical quality attribute (CQA) to be investigated. In case dealing with an impurity a non-inferiority claim is sufficient. Otherwise, checking for equivalence is requested.

2. Define CQAs to be investigated

   • CQAs might differ in their mathematical nature (continuous or binary)
   • Although there are slight differences between the application at EMA and FDA, both agencies do not expect that all quality attributes are identical between the RP and TP. However, prior to the actual analytical biosimilarity comparability assessment, quality attributes should be ranked according to their critical relevance and impact on efficacy and safety. Depending on this ranking, varying rigor in the biosimilarity assessment might be applied. For CQAs that cannot be quantified or have less impact on clinical outcomes, graphical comparison of the raw data is suggested.

3. Define similarity condition

   • Similarity condition is a term used by the recent EMA reflection paper [7] to define an a priori agreement on when two data distributions are to be considered as “similar”, i.e. what is the maximum allowed difference between two underlying distributions. This decision making benefits from a knowledge of the impact such differences could have on clinical outcome. In practise, these impacts are usually not known and risk assessments need to be added to submissions to support the definition of an appropriate similarity condition.
   • It is important to note that EMA stresses to always agree on a similarity condition before a similarity criterion is applied.

4. Definition of a statistical test/ “similarity criterion”

   • The term similarity criterion was introduced in the latest EMA reflection paper and is understood as the concrete instruction in how to use data to make any statements of the a priori agreed similarity condition. In practise, the similarity criterion can be understood as the actual test procedure.
   • Any test employed should show a defined Type I error (error of the agency of wrongly declaring biosimilarity), sometimes also called false positive rate.
   • The latest EMA reflection paper notes that understanding the operating characteristics of each test is important, i.e. the chance of false positive/false negative results. It is expected from the agency that applicants discuss operating characteristics and justify acceptable low chances of false positive conclusions, i.e. a false conclusion for similarity where actually no similarity exists.

5 Conduct experimental study plan and sampling strategy controlling for measurement variability

   • e.g. estimate analytical sample size to account for measurement variability. In case within batch variability (analytical variability) is larger than between batch variability this can be reduced by taking replicates of the sample. In specific cases this even might lower the number of required manufacturing lots for the biosimilar to be produced.

6 Perform equivalence/non-inferiority testing

   • e.g. two one-sided \( t \) (TOST) test and range test.

7 Consideration regarding false positive conclusion and risk mitigation of non-comparability results


\[ \text{Springer} \]
• A usual and expected drawback of current analytical biosimilarity analysis is that for each CQA a test is conducted. This leads to a well-known phenomenon of multiplicity where the false positive rate (Type I) error and the Type II error of the entire test are increased. However, addressing multiplicity is not the focus of this contribution.

This contribution will focus on a novel statistical test for the comparison of quality attributes — also called analytical biosimilarity assessment — between the reference product (RP) and the biosimilar candidate (TP), which is required to achieve licensure [4, 6]. The same comparability exercise is of relevance when evaluating the impact of a change in the manufacturing process [9].

We aim to especially focus on two critical aspects of this workflow: the similarity condition and the similarity criteria. Without having a clear mind about the similarity condition, any formulation of a similarity criterion (realisation of a test) is without meaning. In the past following similarity conditions have been used:

• Previously for equivalence tests on the mean between two distributions, FDA proposed to use 1.5 times the estimated standard deviation as an equivalence acceptance criterion. However, this guideline was withdrawn by FDA.

• Although not mentioning explicitly, a previous publication tailored to the biosimilar comparison task, defined the “equivalence region” if at least the central 99.7% of the TP are within the central 99.7% of RP [10].

• In the area of pre- and post-change comparison, a recent publication used the process capability or out of specification (OOS) rate to set up similarity conditions [11]. Assuming that the specification of the RP equals to 3σ of the RP process, the definition of this similarity condition will be very similar to what we will define in this publication (see the “Results” section). Although this is used frequently in practise, in this case the calculated specification is only a point estimate of the sample mean and standard deviation and therefore associated with sampling error.

In terms of similarity criteria/statistical tests mainly two main approaches were used for claiming biosimilarity for quantitative CQAs in BLA/MMA/NDA fillings in the past:

• The first one is using two one-sided test (TOST) to demonstrate that the difference of mean between biosimilar and innovator is within acceptance criteria. As example, when looking at Mvasi [12], the biosimilar to Avastin from Amgen, similarity of binding to VEGF and anti-proliferation activity of HUVEC cells is assessed using TOST (two one-sided t test).

• The second one is to look at the “population within the population,” which is generally referred to as quality ranges or range tests. One example here is Truxima, the biosimilar to Mabthera from Celltrion [13]. Similarity of quality attribute was claimed, when most of biosimilar lots fell within the range calculated based on 3 times the sample standard deviation of innovators.

**Flaws of Equivalence Tests and Range Tests**

It is important to note that that both equivalence tests and simple range test approaches have flaws and do not comply with all regulatory requirements.

**Equivalence Tests**

Equivalence tests are designed to test null hypothesis that distributional parameters such as means or variances of TP and RP do not depart too much from each other. If this is the case the populations are called “equivalent.” In order to establish what “not too much” means, quantitative equivalence boundaries have been tried to be established [14]. Although there have been some commonly accepted equivalence acceptance criteria proposed (such as $1.5 \times \sigma_{RP}$ for the TOST of difference in means for showing analytical biosimilarity, or 80–125% of the RP for average similarity in evaluating PK data) all of these limits depend on arbitrarily chosen limits. Those limits may be adapted according to the criticality of the CQA to account for residual risk.

The most frequently applied equivalence test of the past in analytical biosimilarity testing is the TOST test [14–16]. This test, for example, examines only the mean difference of RP and TP, and does not take differences in variance into account. In a rather theoretical but extreme case of a very large number of TP and RP lots, it is possible to claim similarity for a biosimilar product with a large difference of variance but a small difference of mean to innovators [10]. To be precise, in this case only the mean difference needs to be smaller than the equivalence margin. Since this test is not present any more in the updated FDA guideline we will not go into detail of this test here and refer to literature elsewhere which even proposes alternative tests with better power [17].

There also exist equivalence tests for variances ($F$-tests as described in [18]). However, a separate equivalence test on mean and variance neglects the interplay between both to still receive acceptable product; e.g. one could think of a biosimilar which has some (even larger) mean difference to the RP. However, due to its small variability it is still producing a lot of acceptable product. Also vice versa, a biosimilar candidate that shows less or no mean difference to the RP may accept a bit higher variability (compare Fig. 2).
elements are not taken into account when performing separate equivalence tests on mean and variance.

**Range Tests**

A quality range approach as defined by FDA is a biosimilarity assessment method that takes into account both sample mean and sample variance to overcome that limitation of the TOST test. Moreover, they take into account the combination between both distributional parameters to reach acceptable product. The rationale is straightforward: if most of TP batches are within RP population, it is likely that TP distribution is within the range of the RP.

A frequently applied realisation of this concept is to claim biosimilarity when a sufficiently large percentage of TP batches (e.g. 90%) falls within the acceptable range represented by the RP population. We want to note that in case of less than 10 TP batches usually all batches need to be within a given acceptable range. Several methods have been used to define the acceptable range, such as Min–Max of all RP batches (so called Min–Max test), three standard deviation (so called 3SD test), or tolerance interval. Amongst them, Min–Max has been considered as the range with less regulatory risk and more manufacture risk [10]. On the contrary, as described in Fig. 2 of [10], tolerance interval is the range with more regulatory risk and less manufacture risk as they show the highest false acceptance rate at typical sample size of \( n = 10 \). For these sample sizes, according to Fig. 2 of [10], the three standard deviation test [16] is the one of three methods that balance between regulatory and manufacture risks as it shows comparably low false acceptance rates and low false rejection rates and it was similar to the quality range method recommended by FDA [6]. For all of these simple range tests, such as the 3SD test or the Min–Max test, it is generally easier to claim biosimilarity with less TP batches, which may discourage the sponsor to increase the manufacture batches [10]. Statistically speaking, the false positive error (Type I error) of range test — the chance of concluding biosimilarity although there is none — is a function of the sample size. This is not the preferred behaviour of a proper statistical test which usually keeps the Type I error at a constant, pre-defined level, usually at 5%. Regulatory agencies have taken note of those flaws and urge for a biosimilarity test that keeps the Type I error, which equals to the regulatory risk, independent of the sample size. Since this is known to agencies there is a need to develop a test that controls for the regulatory risk. In some cases the range test is formulated to pass when a large fraction or all of TP fall within \( \bar{x}_{RP} \pm C \times s_{RP} \), where \( \bar{x} \) and \( s \) are the sample mean and standard deviation, respectively. Then \( C \) can be defined as a function of the similarity condition, \( n_{RP} \) and \( n_{TP} \) to assure a defined alpha level of the test (Type I error) [11]. However, there are certain drawbacks of such a test. For ease of comparability to nomenclature of the previous publication let us define \( K_1 = \frac{|\mu - \mu_{RP}|}{\sigma_{RP}} \) and \( K_2 = \frac{C}{\sigma_{RP}} \), where \( \mu \) and \( \sigma \) are the true population mean and standard deviation.

Figure 1 shows the Type I error as a function of \( C \), \( K_1 \), and \( K_2 \) at a sample size of \( n_{TP} = n_{RP} = 20 \). \( K_1 \) and \( K_2 \) have been chosen in a way where 99% (as a synonym for 3SD) of the TP are within 99% of the RP. This equals the decision boundary between biosimilar and non-biosimilar region as defined in Fig. 2. We see in Fig. 1 that, in order to achieve a 5% Type I error, we need to adjust the levels \( C \) as a function of the levels of \( K_1 \) and \( K_2 \). However, an a priori knowledge of \( K_1 \) and \( K_2 \) is not available in practice; we only have sample estimates. This contribution also aims at investigating the impact of different levels of \( C \) (all originating from \( K_1 \) and \( K_2 \) values of the same similarity condition) on the power and Type I error.

**Requirements to a Novel Test for Showing Analytical Biosimilarity** In this contribution, we want to establish a novel statistical test that reduces the abovementioned flaws and is compliant with current regulatory requirements. Specifically, we will focus on three criteria to accept such a new test:

- The test should have an easy-to-define and clearly formulated null hypothesis based upon a similarity condition to be tested; i.e. the test is designed to rejects the null hypothesis of being not biosimilar. This is currently not fulfilled with simple range test such as the 3SD or Min–Max test.
- The test simultaneously checks for the underlying population and not only single characteristics such as the mean and the variance between the TP and RP. This is currently not achieved by simple equivalence tests such as the TOST test.
- The operation characteristics should be easy to understand and the Type I error (agencies risk) should be controlled along the entire similarity condition and independent of the used sample size of TP and RP. This is of outmost importance and currently not achieved by simple quality range test.

For CQAs that do not pass biosimilarity tests as mentioned above, an extensive characterisation and investigation needs to be performed to understand their potential impact to potency and safety, e.g. in clinical trials. Both the analytical analyses and clinical trial data need to be provided by the sponsor and the agency will evaluate biosimilarity
Material and Methods

Simulation Study to Estimate the Operating Characteristics of Analytical Biosimilarity Tests

To compare the bootstrapping method described in the “Results” section with other analytical similarity tests for biosimilar assessment, we adopted a tool to assess the properties of new statistical tests and their applicability for showing analytical similarity developed by Stangler and Schiestl [10]. As they proposed, we conducted a simulation study where samples for a specific number of TP and RP batches were generated from normal distributions with a variety of possible combinations of $\mu_{TP}$, $\mu_{RP}$, $\sigma_{TP}$, and $\sigma_{RP}$ (known distributional parameters). Only a simulation study with known ground truth serves as a reliable comparator between the tests. For each of these combinations of the distributional parameters the simple Min–Max test, the 3SD test, the TOST test, and the above-described
bootstrapping test were conducted. The Min–Max test and the 3SD test were conducted as defined in the “Introduction” section. For the TOST test the equivalence acceptance criterion was set to 1.5 \times \sigma_{RP}^-. For further description of the tests we refer elsewhere [10]. Each combination of distributional parameters was tested for a large number of repetitions and the rate of acceptance of biosimilarity was recorded for each of the tests and combinations. The procedure of this simulation study is listed below:

1. Select one combinations of \( \mu_{TP}, \mu_{RP}, \sigma_{TP}, \) and \( \sigma_{RP} \) (known population parameters) and build two normal distributions \( N(\mu_{TP}, \sigma_{TP}^2) \) and \( N(\mu_{RP}, \sigma_{RP}^2) \).
2. Randomly sample the specific number of TP and RP batches from normal distributions \( N(\mu_{TP}, \sigma_{TP}^2) \) and \( N(\mu_{RP}, \sigma_{RP}^2) \), respectively.
3. Examine biosimilarity by applying different tests such as the Min–Max test, the 3SD test, the TOST test, and the bootstrapping test.
4. Repeat steps 2 and 3 for a large number of times (e.g. 1000) and count how often the specific test can claim biosimilarity. The chance of passing is calculated by dividing the successful counts of step 3 by the total number of tries (e.g. 1000).

In order to visualise the results, for each combination, the absolute difference in means per standard deviation of the reference \( \frac{|\mu_{TP} - \mu_{RP}|}{\sigma_{RP}} \) as well as the ratio of standard deviations \( \frac{\sigma_{TP}}{\sigma_{RP}} \) was plotted over the acceptance rate (shown in Figs. 4 and 5). In order to generate these plots a grid size of 10 grid points in each dimensions \( \frac{|\mu_{TP} - \mu_{RP}|}{\sigma_{RP}} \) and \( \frac{\sigma_{TP}}{\sigma_{RP}} \) was chosen. This normalisation has been already applied previously in literature and shown to be useful for the purpose of displaying operating characteristics (10).

Additionally, the here employed null hypothesis, with \( \sim 99\% (\mu_{TP} \pm 3 \times \sigma_{TP}) \) of the TP population is within \( \sim 99\% (\mu_{RP} \pm 3 \times \sigma_{RP}) \) of the RP population, can be represented by the area under orange decision lines in each subplot of Figs. 4 and 5. To fulfill this biosimilarity criteria \( \sigma_{TP} \) needs to be smaller than \( \sigma_{RP} \), when \( |\mu_{TP} - \mu_{RP}| > 0 \). To be precise, following inequality needs to be met: \( \frac{\sigma_{TP}}{\sigma_{RP}} \geq 1 - \frac{|\mu_{TP} - \mu_{RP}|}{3 \times \sigma_{RP}} \). Thereby, the decision boundary line can be depicted as the linear equation of \( \frac{|\mu_{TP} - \mu_{RP}|}{3 \times \sigma_{RP}} + \frac{\sigma_{TP}}{\sigma_{RP}} = 1 \) (see orange line in Figs. 2, 4, and 5).

The mean false acceptance rate (Type I error) was calculated by taking the mean acceptance rate at the decision boundary (grid points along the boundary line). The mean acceptance rate within the region below the orange line is the power (1 – Type II error or 1 – false rejection rate) of the test. Ideally, a perfect test would be able to discriminate exactly at this line, having an acceptance rate of 1 for all combinations of \( \frac{|\mu_{TP} - \mu_{RP}|}{\sigma_{RP}} \) and \( \frac{\sigma_{TP}}{\sigma_{RP}} \) below that line (green biosimilar area in Fig. 2) and an acceptance rate of 0 above that line (non-biosimilar area in Fig. 2). However, this would require an infinite amount of TP and RP batches. For typical statistical tests, we would expect that at the border line the false acceptance rate (= Type I error) equals a pre-defined significance limit, usually set to 0.05.

### Results

A Novel Bootstrapping Test for Showing Analytical Biosimilarity

**Definition of the Similarity Condition/Null Hypothesis for Biosimilarity**

We want to propose an easy-to-implement bootstrapping method that controls the Type I error for range tests and the other acceptance criteria listed in the “Introduction” section. Prior to conducting the test, we define a similarity condition — or statistically speaking null hypothesis. Inspired by Fig. 2, the major idea behind the similarity condition employed in this contribution is that a specific fraction of TP batches (for this contribution we have chosen \( \sim 99\% \) of a normal distribution) are within a specific range of the RP batches (also \( \sim 99\% \) of the distribution of RP). This has been also done by other authors in the past [10] and is linked to the idea that the drug product specifications should be at least cover 99\% of the RP. Therefore, by fitting the TP within those boundaries the biosimilar should show feasible quality also in terms of drug substance specification of the innovator, which is usually not known to the biosimilar sponsor. Mathematically this specific null hypothesis can be formulated by \( \frac{\sigma_{TP}}{\sigma_{RP}} < 1 - \frac{\mu_{TP} - \mu_{RP}}{3 \times \sigma_{RP}} \) (see further details in the “Material and Methods” section). Of course this selection of the null hypothesis can be adapted according to criticality of the CQA.

As indicated in the “Introduction” section, another contribution [11] used a defined OOS rate as a similarity condition. In case the specification for the TP process equals \( \mu_{RP} \pm 3 \times \sigma_{RP} \) and applying an acceptable OOS rate of 0.3\% (100–99.73\%) the approaches are quite similar as indicated in Figure S1 of the supporting information. The difference is that the definition of the similarity condition using the OOS rate leads to a non-linear decision boundary whereas the definition of having 99\% of TP within 99\% of RP leads to a linear decision boundary. For practical purposes both similarity conditions can be considered very similar. Per definition the similarity condition is a priori agreement on when two data distributions are to be
considered as “similar.” Hence the definition of the any similarity condition will be subjective.

**Implementation of the Bootstrapping Test**

The null hypothesis is that product is not biosimilar, i.e., 99% of TP is not within 99% of RP, since we want to reject that and accept for biosimilarity. Therefore,

\[
H_0 : \frac{\sigma_{TP}}{\sigma_{RP}} - 1 + \frac{|\mu_{TP} - \mu_{RP}|}{3 \times \sigma_{RP}} \geq 0
\]

More general:

\[
H_0 : \frac{\sigma_{TP}}{\sigma_{RP}} - \Phi^{-1}(p_{RP}) + \frac{|\mu_{TP} - \mu_{RP}|}{\Phi^{-1}(p_{TP}) \times \sigma_{RP}} \geq 0
\]

where \( \Phi^{-1}(p_{RP}) \) and \( \Phi^{-1}(p_{TP}) \) are the inverse of the cumulative standard normal distribution given the required probabilities of RP and TP, respectively.

The test is implemented using four steps:

1. Start with a given sample of TP and RP batches
2. Resample with replacement the exact same number of TP and RP batches. This is a very typical resampling procedure for bootstrapping approaches [19].
3. Repeat step 2 for a large number of times (e.g. 1000) and count how often the null hypothesis can be rejected.
4. A chance of failing is calculated by dividing the number of rejections of step 3 by the total number of tries (e.g. here 1000). If this chance, which can also been seen as a bootstrapped version of the \( p \)-value, is below a predefined significance level (e.g. 5%) we reject the null hypothesis and conclude for biosimilarity.

Figure 3 shows an exemplary result of the proposed test. In this case the biosimilarity can be concluded with a significance of 99.4% (\( p \)-value of 0.006). The isolines of the contour plot illustrate the likelihood of the bootstrapped samples but should only serve as a visual tool to illustrate the distribution. The estimate of the distribution was made using a Gaussian kernel function with the Python package (sklearn.neighbors.KernelDensity(kernel = 'gaussian', bandwidth=0.3), sklearn version 0.20.2). The outcome of the test (rejection or not) as well as the \( p \)-value is only calculated by the number of green dots outside the equivalence region.

**Comparison of the Bootstrapping Test With Other Analytical Similarity Tests**

The results of the simulation study can be seen in Figs. 4 and 5, for \( n_{TP} = n_{RP} = 8 \) and \( n_{TP} = n_{RP} = 20 \), respectively.

As we can see in the subplots of the Min–Max tests, the 3SD, and the TOST test of Fig. 4, the acceptance rate at and above the orange decision line is far higher than 5%.
However, in an ideal scenario we do not expect the acceptance rate to be higher than 5% at the orange decision boundary. Moreover, even stronger deviation from biosimilarity going in the direction of the right top corner of each subplot (compare “non-biosimilar” area of Fig. 2) still leads to high acceptance rates for the Min–Max tests, the 3SD, and the TOST test. In more detail, at the orange decision line we get the false acceptance rate of 10–25% (bright blue area) for the Min–Max tests and approximately 50–80% (yellow and red area) for the 3SD test. The TOST test is known to control the Type I error in the direction of mean difference and therefore the acceptance rate at $\frac{|\mu_{RP} - \mu_{TP}|}{\sigma_{TP}} = 1.5$ equals 0.05. However, the TOST does not explicitly test for differences in the ratio of standard deviations. Therefore, in specific situations of small mean differences of, e.g. 0.5 and large $\frac{\sigma_{RP}}{\sigma_{TP}}$ of, e.g. 1.75, TOST test is even more willing to wrongly accept for biosimilarity than the Min–Max test.

When comparing those different tests to each other one needs to take their different characteristics into account originating from different null hypothesis. For example, it is obvious that due to the nature of the TOST test as being an equivalence tests on the means it will not control for any differences of variances between the RP and the TP. However, all the tests discussed in this contribution are or have been widely used in industry to assess analytical biosimilarity and their impact on patient safety and efficacy needs to be understood.

In the case of increasing the sample size of TP and RP to 20 each, the results are presented in the same way in Fig. 5. For the TOST test setting we can see that wrongly accepting for biosimilarity increases for small mean difference and large SD ratios. In such a scenario a majority of the manufactured TP batches is not within a range where ~ 99% of the RP can be expected. The 3SD test stays at a similar acceptance level at the orange border line compared to Fig. 4 with $n = 8$. Nevertheless a reduction in acceptance rate in the non-similarity region (above the orange line) can be observed since the bright blue area is decreased for the case with $n = 20$. This is contradictory to an expected behaviour of a statistical test, which should favour the manufacturer’s chances to pass in case of producing a truly biosimilar and increasing sample size. However, the 3SD test seems to work in exactly the different way, where low sample sizes are rewarded. On the other hand, the proposed bootstrapping test controls the Type I error at the equivalence border (orange line) at 5% (see subplot “BS” in Figs. 4 and 5 and Table I). Moreover, the chances of correctly accepting for similarity (here in case 99% of TP is within 99% of RP population) increase when sample size is increased. This can be seen as the red area in subplot BS (bootstrapping) of Fig. 5 grows compared to subplot BS of Fig. 4. Of course, when moving from the frequently used 3SD or Min–max test to the proposed bootstrapping test an overall reduced chance of passing the similarity assessment can be expected. This can be seen in Figs. 4 and 5 since the overall acceptance rate (e.g. mean of the entire surfaces shown) is always lower for the bootstrapping test compared to the 3SD or Min–Max test. Nevertheless, the bootstrapping test is the only one which controls properly the Type I error, which can be expected to be a regulatory requirement in the future.

The numerical results of the mean false acceptance rate at the decision boundary of the null hypothesis, as well as the mean true acceptance rate within the biosimilar area, can be seen in Table I. Here it can be seen that the BS test leads to an approximate-numerically 5% Type I error regardless of the used sample size for RP and TP. The mean true acceptance rate (power of the test) for the BS test increases with increasing sample size. These are the properties we require from a solid statistical test. In contrast, the 3SD test and the
Min–Max test shows extremely high false acceptance rates at the decision boundary. The 3SD test even have increased Type I error at low sample sizes. This is of course a very undesirable property that rewards companies of testing less lots than more.

In addition to the comparison of the bootstrapping test to the TOST, Min–Max, and 3SD tests, it is also compared to a recently developed range test [11] that aims at controlling the Type I error. Similarity is concluded when all or a specific fraction of TP runs fall within $x_{\text{RP}} \pm C \times s_{\text{RP}}$. $C$ is calculated based on an a priori selection of $K_1$ (equivalent to $\frac{\mu_{\text{TP}} - \mu_{\text{RP}}}{\sigma_{\text{TP}}}$ in this contribution) and $K_2$ ($\frac{\sigma_{\text{TP}}}{\sigma_{\text{RP}}}$) along the similarity condition [11]. Table II shows different values for $C$ given $K_1$ and $K_2$ values that obey the similarity condition. Figure 6 shows the acceptance rates of the range test of [11] and the bootstrapping test with a sample size of TP and RP equal to 20. The green dots in the subplots of the range test indicate the chosen values of $K_1$ and $K_2$. Visually the acceptance rates of the range test of [11] with the different selections of $K_1$ and $K_2$ look very different from each other specifically when choosing a $C$ value of 2.3. Moreover, it becomes clear that the Type I error is properly controlled at the green dots but not on other points of the decision boundary. In contrast the presented bootstrapping test controls the Type I error over the entire decision boundary more evenly. This becomes also clear when looking at the numerical values of the acceptance rate at 10 distinct points of the decision boundary, as shown in Table III. This table can be seen as a numerical excerpt of the acceptance rates along the orange decision boundary of Fig. 6. Only at the specific values of $K_1$ ($\frac{\mu_{\text{TP}} - \mu_{\text{RP}}}{\sigma_{\text{TP}}}$) and $K_2$ ($\frac{\sigma_{\text{TP}}}{\sigma_{\text{RP}}}$) that were used for calculating the $C$ value the nominal value of 5% Type I error can be achieved (see italic entries in Table III). Specifically the choice of $K_1 = 3$ and $K_2 = 0$, which leads to a $C$ value of 2.3, seems to be very unfavourable to reach correct Type I errors for any other combination of $K_1$ and $K_2$ of the decision boundary.

Table 1 Mean True Acceptance Rate and Mean False Acceptance Rate of 3SD, BS, and Min–Max Tests. For Mean True Acceptance Rate for Each Test, the Mean of Individual Acceptance Rates of Combinations of $\mu_{\text{TP}}$, $\mu_{\text{RP}}$, $\sigma_{\text{TP}}$, and $\sigma_{\text{RP}}$ Within the Biosimilar Area (Compare Green Shaded Area of Fig. 2) Was Calculated. For Mean False Acceptance Rate for Each Test, the Mean of Individual Acceptance Rates of Combinations of $\mu_{\text{TP}}$, $\mu_{\text{RP}}$, $\sigma_{\text{TP}}$, and $\sigma_{\text{RP}}$ at the Decision Boundary (Compare Orange Line in Figs. 2, 4, and 5) Was Calculated.

| $n_{\text{RP}} = n_{\text{TP}}$ | Mean false acceptance rate at decision boundary (Type I error) | Mean true acceptance rate (power of the test) |
|---|---|---|
| | 3SD | Min–Max | BS | 3SD | Min–Max | BS |
| 8 | 0.75 | 0.14 | 0.03 | 0.89 | 0.49 | 0.32 |
| 15 | 0.75 | 0.15 | 0.03 | 0.94 | 0.57 | 0.44 |
| 30 | 0.79 | 0.19 | 0.04 | 0.96 | 0.66 | 0.61 |
| 100 | 0.76 | 0.33 | 0.05 | 0.97 | 0.8 | 0.76 |
| 1000 | 0.29 | 0.62 | 0.05 | 0.92 | 0.95 | 0.93 |

Fig. 5 Acceptance rate of 4 different tests (Min–Max, 3SD, TOST, and the presented bootstrapping (BS) method) over a variety of combinations of mean differences normalised by the standard deviation of the reference ($\frac{\mu_{\text{TP}} - \mu_{\text{RP}}}{\sigma_{\text{RP}}}$) and ratio of standard deviations ($\frac{\sigma_{\text{TP}}}{\sigma_{\text{RP}}}$). Sample size for TP and RP equals 20. Orange lines in each subplot show the border where ~99% (3SD) of the TP population is within ~99% (3SD) of the RP population. Significance level of $\alpha = 0.05$ was employed for the bootstrapping test. Colour code indicates the acceptance rate of the test.
Also choosing a $K_1 = 0$ and $K_2 = 1$ will lead to very low Type I errors, even going to 0, when higher values of $K_1$ and lower values of $K_2$ are actually present in the true underlying data set, e.g. $K_1 = 1.667$ and $K_2 = 0.444$. Although the bootstrapping test also shows variation in the Type I error along the decision boundary — specifically at the edges, i.e. at $K_1=0/\kappa_2=1$ and $K_1=3/\kappa_2=0$ — the desired Type I error of 5% is more evenly reached.

As shown in Table IV, when comparing the mean power of the range test from [11] to the bootstrapping test of this contribution it turns out that for all sample sizes, the bootstrapping tests yield higher power values, whilst still keeping the Type I error closer to the desired level of 5%. There are situations where the mean power of the range tests from [11] yields better power, i.e. when going to $K_1 = 3$ and $K_2 = 0$ for calculating the $C$ value, but in those cases the mean Type I error by far exceeds the 5% margin. Please note that for all simulations of the range tests from [11] in Table IV the value of $C$ was adapted for each sample size according to Table II. In cases when choosing $K_1 = 1$, $K_2 = 0.66$ for the range test and looking at sample size equal or larger than 20 the bootstrapping test leads to 10% more power. Please note that all reported power values in Table IV are only mean power values that also include conditions that are very close to the decision boundary. In case the TP distribution is more narrow than the RP distribution or very close to the mean of the RP the power values are much higher as indicated in the BS subplot of Fig. 6.

However, for small sample sizes the mean power even for the bootstrapping test is still low and in practical cases passing the test might be difficult. Hence, either a higher Type I error of, e.g. 0.1 might be accepted by regulators or the similarity condition might be changed. Here we
Table III  Acceptance Rates for Sample Size of TP and RP Equal to 20 for the Bootstrapping Test (BS) and the Range Test of (11) With Different Levels of C Originating From Selection of \( K_1 \) and \( K_2 \) Obeying the Decision Boundary. Italic Entries Indicate the Acceptance Rate at the Value of the Decision Boundary That Was Also Selected to Calculate C

| \( \frac{n_{RP} + n_{TP}}{\sigma_{TP}} \) | \( \frac{n_{RP}}{\sigma_{RP}} \) | BS | C = 1.42 | C = 1.49 | C = 2.3 |
|---|---|---|---|---|---|
| (\( K_1 = 0 \), \( K_2 = 0.667 \)) | (\( K_1 = 1 \), \( K_2 = 1 \)) | (\( K_1 = 3 \), \( K_2 = 0 \)) |
| 0.00 | 1.00 | 0.01 | 0.05 | 0.07 | 0.54 |
| 0.33 | 0.89 | 0.05 | 0.08 | 0.10 | 0.63 |
| 0.67 | 0.78 | 0.04 | 0.07 | 0.09 | 0.62 |
| 1.00 | 0.67 | 0.07 | 0.05 | 0.06 | 0.55 |
| 1.33 | 0.56 | 0.04 | 0.01 | 0.01 | 0.38 |
| 1.67 | 0.44 | 0.03 | 0.00 | 0.00 | 0.32 |
| 2.00 | 0.33 | 0.03 | 0.00 | 0.00 | 0.21 |
| 2.33 | 0.22 | 0.03 | 0.00 | 0.00 | 0.14 |
| 2.67 | 0.11 | 0.02 | 0.00 | 0.00 | 0.08 |
| 3.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.06 |


demonstrate how a change in the similarity condition to a claim where 90% of the TP should be within 99% of the RP can be implemented using the bootstrapping test. For that the more general null hypothesis can be applied by setting \( p_{TP} \) to 0.9 and \( p_{RP} \) to 0.99.

Figure 7 shows the acceptance rate of the bootstrapping test under this new similarity condition with sample size of TP and RP equal to 20. The decision boundary is now shifted compared to the initial decision boundary as depicted in, e.g., Figure 5. Compared to the initial acceptance rates of the bootstrapping test in Fig. 5, the acceptance rates in Fig. 7 increased substantially and power values in the original similarity condition of >50% and in most cases >80% can be achieved.

Table IV  Mean True Acceptance Rate and Mean False Acceptance Rate of Three Versions of the Range Tests From (11), Each Calibrated With Different Choices of \( K_1 \) and \( K_2 \). As Well As The Bootstrapping (BS) Test. For Mean True Acceptance Rate for Each Test, the Mean of Individual Acceptance Rates of Combinations of \( \mu_{TP} \), \( \sigma_{TP} \), and \( \sigma_{RP} \) Within the Biosimilar Area (Compare Green Shaded Area of Fig. 2) Was Calculated. For Mean False Acceptance Rate for Each Test, the Mean of Individual Acceptance Rates of Combinations of \( \mu_{TP} \), \( \mu_{RP} \), \( \sigma_{TP} \), and \( \sigma_{RP} \) at the Decision Boundary (Compare Orange Line in Figs. 2, 4, and 5) Was Calculated

| Mean false acceptance rate at decision boundary (Type I error) | Mean true acceptance rate (power of the test) |
|---|---|
| Range test from [11] | BS |
| \( n_{RP} = n_{TP} \) | \( n_{RP} = n_{TP} \) |
| \( K_1 = 0 \) | \( K_1 = 1 \) | \( K_1 = 3 \) |
| \( K_2 = 1 \) | \( K_2 = 0.67 \) | \( K_2 = 0 \) |
| \( K_1 = 0 \) | \( K_1 = 1 \) | \( K_1 = 3 \) |
| \( K_2 = 1 \) | \( K_2 = 0.67 \) | \( K_2 = 0 \) |

Discussion

Due to the recent withdrawal of the previous draft FDA guideline in 2018 and the latest emergence of the EMA reflection paper on statistical approaches for analytical similarity assessment [7] the need for critical discussion and development of new regulatory compliant statistical procedures for analytical similarity assessment and comparability testing is a young scientific area. Therefore, to the best of our knowledge, only two studies have dealt with the current regulatory limitations of quality range tests in analytical biosimilarity testing [10, 11]. Within their publication Stangler and Schiestl [10] have critically investigated the unfavourable property of existing equivalence and range tests for analytical similarity testing, such as the TOST test, 3SD test, and Min–Max test. They have presented a tool to investigate the properties and applicability of different analytical similarity tests by calculating operating characteristics, i.e. the mean false acceptance rate and the mean false rejection rate. The same methodology was used in our publication with some additional improvements. The mean false rejection rate of their work is equivalent to \( 1 - \text{power calculated in our contribution} \). The mean false acceptance rate was calculated in their contribution by taking the mean over all grid points of the non-equivalence region. Therefore, the value of the mean false acceptance rate is a function of the chosen arbitrary size of the non-equivalence region. In order to reduce subjectivity the mean false acceptance rate in this contribution was only calculated at grid points that equal the similarity decision boundary (compare orange lines in Figs. 2, 4, and 5). This has the advantage that the value now expresses the actual Type I error of the test and thereby more statistical meaning in the interpretation of the value.
The AAPS Journal (2022) 24:112

The observation in our contribution that the false acceptance rate (here Type I error) is a function of sample size of TOST tests and simple range tests and generally decreases with increasing sample size has also been made by Stangler and Schiestl [10]. Herein, they state the need to develop tests that are able to control the Type I error independent of the sample size, in which succeeded in this publication. Here it should be noted that all the tests, e.g. TOST test or the proposed bootstrapping test, have different null hypothesis and when comparing these tests that needs to be taken into account. However, a comparison is still useful as all of these tests presented in this contribution are used in the context of analytical biosimilarity assessment and possible impact releasing a biosimilar to the marketing including its effect on patient safety and efficacy needs to be understood.

In a second contribution Burdick [11] developed a range test that controls the Type I error for specific selections of mean differences normalised by the standard deviation of the reference ($\frac{\mu_T - \mu_R}{\sigma_R}$) and ratio of standard deviations ($\frac{\sigma_T}{\sigma_R}$). The similarity condition used by Burdick [11] is based on a constant OOS rate and shown to be very similar to the one used in this contribution (Figure S1 in supplementary information). One could also adapt the bootstrapping test to accept the similarity condition proposed by Burdick. Thereby, the decision boundary would no longer be linear.

![Figure 7](image-url)  
Fig. 7 Acceptance rate of the bootstrapping (BS) method. The acceptance rates are shown over a variety of combinations of mean differences normalised by the standard deviation of the reference ($\frac{\mu_T - \mu_R}{\sigma_R}$) and ratio of standard deviations ($\frac{\sigma_T}{\sigma_R}$). Sample size for TP and RP equals 20. Orange lines in each subplot show the border where 90% of the TP population is within 99% of the RP population. Significance level of $\alpha = 0.05$ was employed. Colour code indicates the acceptance rate of the test.
but non-linear. Again, the selection is a somewhat subjective and general consideration of what kind of two underlying distributions can be considered as similar/non-similar.

The presented novel bootstrapping test for analytical similarity testing for biosimilar assessment has shown required qualities in terms of a controlled Type I error (agency’s risk) to falsely declare a non-biosimilar being biosimilar) and increasing power with increasing sample size (see Figs. 4 and 5 for visual comparison to other analytical similarity tests as well as Table I for numerical values). Nevertheless, the absolute chance to pass the novel bootstrapping test is drastically reduced at low sample sizes (< 30) compared to other flawed tests such as 3SD or the Min–Max test (see Table I). Due to lowered chances of passing the test, manufacturers will need to adapt their strategy during biosimilar process development. One possibility to increase the success rate with the bootstrapping test is to make sure to produce as closely as possible at the innovator mean and with a lowest possible variation in the process. This equals the lower left region of all subplots in Figs. 4 and 5. Thereby, power levels of the novel bootstrapping test above 80% can be reached even at low sample size cases (see Fig. 4). Another consequence could be to increase sample size for TP and RP; however, this is cost-intensive. A third option could be to relax the Type I error or the null hypothesis of defining biosimilarity. As demonstrated in Fig. 7, a relaxed similarity condition of 90% of the TP to be within 99% of the RP, as also proposed by others [10, 11], can be applied and even higher power values reached. Of course, these relaxations need to be well aligned with the criticality of the studied CQAs.

### Conclusion

In this contribution, we have introduced a novel bootstrapping test for demonstrating analytical biosimilarity that is perceived compliant with regulatory expectations. It accounts for the following acceptance criteria:

- The test has a clearly formulated null hypothesis to be tested; i.e. a specific fraction (e.g. 99%) of the TP should be within a specific fraction (e.g. 99%) of the RP. This has been already proposed in a similar way by Stangler and Schiestl [10] and relates to the concept of similarity condition by EMA’s recent reflection paper. This similarity condition is easy to establish and can be modified according to CQA criticality.
- The test checks for differences in means and variance between the TP and RP, simultaneously.
- The Type I error (agency’s risk) is controlled at a predefined limit (e.g. 5%) on the entire decision boundary of the similarity condition and independent of the used sample size of TP and RP.

The novelty is that this test is the only quality range test so far that controls for the Type I error evenly on the similarity condition for varying sample sizes of RP and TP. This is also the reasons why this test is superior compared to other frequently employed tests such as the 3SD test, the Min–Max tests, or other modified range tests [11]. Simple range tests, such as the 3SD test, show very large Type I errors (high chances of passing the test although the product is not biosimilar), which at the same time leads to high chances of passing test in case the product is biosimilar. The control of the Type I error of the presented bootstrapping test comes at the cost of also lowering the power of the test (chance of passing in case the product is biosimilar). This will lead to an increased focus of the manufacturer to produce highly similar product in terms of low mean difference and low production variability or at a purposeful design of relaxed similarity conditions.

By simultaneously taking into account differences in means and variance the test is also superior to classical equivalence testing that either account for differences in means (TOST test) or differences in variances (F tests).

The proposed test can be used for analytical biosimilarity testing but also to compare process performance and quality after a manufacturing change in case two (unknown) populations need to be compared.

Since the proposed bootstrapping test is designed to test CQAs individually, multiplicity originating from conducting multiple of the proposed tests is not controlled and might be issue that needs to be investigated and addressed in the future.

### Supplementary Information

The online version contains supplementary material available at [https://doi.org/10.1208/s12248-022-00749-3](https://doi.org/10.1208/s12248-022-00749-3).

### Acknowledgements

Many thanks to Prof. Christoph Herwig (Technical University of Vienna), Shin-Kai Lin (JHL Biotech), and Mahalia Corrales (JHL Biotech) for the fruitful discussion and review of the manuscript.

### Author Contribution

Thomas Zahel designed the new analytical similarity test for biosimilar assessment, performed the simulation studies, and wrote the manuscript.

### Declarations

#### Conflict of Interest

The author declares no competing interests.

#### Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).
References

1. Santos SB, Sousa Lobo JM, Silva AC. Biosimilar medicines used for cancer therapy in Europe: a review. Drug Discovery Today. 2019;24(1):293–9.
2. Nupur N, Joshi S, Gulliarme D, Rathore AS. Analytical similarity assessment of biosimilars: global regulatory landscape, recent studies and major advancements in orthogonal platforms. Front Bioeng Biotechnol. 2022;9(10): 832059.
3. FDA. Scientific considerations in demonstrating biosimilarity to a reference product guidance for industry. 2015;27.
4. EMA. Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues (revision 1). 2014.
5. Wang J, Chow SC. On the regulatory approval pathway of biosimilar products. Pharmaceuticals. 2012;5(4):353–68.
6. FDA. Development of therapeutic protein biosimilars: comparative analytical assessment and other quality-related considerations. Center Biologics Eval Res. 2019.
7. EMA. Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development. 2021;21.
8. Tsong Y, Dong X, Shen M. Development of statistical methods for analytical similarity assessment. J Biopharm Stat. 2017;27(2):197–205.
9. ICH. ICH Q5E: comparability of biotechnological/biological products subject to changes in their manufacturing process. Step 5. 2005.
10. Stangler T, Schiestl M. Similarity assessment of quality attributes of biological medicines: the calculation of operating characteristics to compare different statistical approaches. AAPS Open. 2019;5(1):4.
11. Burdick RK. Statistical considerations for comparative assessment of quality attributes. Statistics in Biopharmaceutical Research. 2020;8:1–6.
12. Seo N, Polozova A, Zhang M, Yates Z, Cao S, Li H, et al. Analytical and functional similarity of Amgen biosimilar ABP 215 to bevacizumab. mAbs. 2018;10(4):678–91.
13. Lee KH, Lee J, Bae JS, Kim YJ, Kang HA, Kim SH, et al. Analytical similarity assessment of rituximab biosimilar CT-P10 to reference medicinal product. mAbs. 2018;10(3):380–96.
14. Chung Chow S. On assessment of analytical similarity in biosimilar studies. Drug Designing: Open Access [Internet]. 2014 [cited 2016 Dec 18];03(03). Available from: http://www.omicsgroup.org/journals/assessment-of-analytical-similarity-in-biosimilar-studies-2169-0138.1000e124.php?aid=35519. Accessed 1.09.2022.
15. Chow SC. Analytical similarity assessment in biosimilar product development [Internet]. 1st ed. Chapman and Hall/CRC; 2018 [cited 2019 Jun 1]. Available from: https://www.taylorfrancis.com/books/9781351339469. Accessed 1.09.2022.
16. Burdick R, Coffey T, Gutka H, Gratzl G, Conlon HD, Huang CT, et al. Statistical approaches to assess biosimilarity from analytical data. AAPS J. 2017;19(1):4–14.
17. Berger RL, Hsu JC, et al. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Stat Sci. 1996;11(4):283–319.
18. Zhang N, Chi E. Statistical considerations for the development of biosimilar products. GaBI J. 2014;3(1):21–5.
19. Efron B. The jackknife, the bootstrap, and other resampling plans. Vol. 38. Siam; 1982. https://doi.org/10.1137/1.9781611970319.fm.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.