The AAPS Journal (2019) 21: 7
DOI: 10.1208/s12248-018-0275-9

Research Article

The Assessment of Quality Attributes for Biosimilars: a Statistical Perspective on Current Practice and a Proposal

Johanna Mielke,[1] Franz Innerbichler,[2] Martin Schiestl,[2] Nicolas M. Ballarini,[3] and Byron Jones[1,4]

Received 14 August 2018; accepted 24 October 2018; published online 27 November 2018

Abstract. Establishing comparability of the originator and its biosimilar at the structural and functional level, by analyzing so-called quality attributes, is an important step in biosimilar development. The statistical assessment of quality attributes is currently in the focus of attention because both the FDA and the EMA are working on regulatory documents for advising companies on the use of statistical approaches for strengthening their comparability claim. In this paper, we first discuss "comparable" and "not comparable" settings and propose a shift away from the usual comparison of the mean values: we argue that two products can be considered comparable if the range of the originator fully covers the range of the biosimilar. We then introduce a novel statistical testing procedure (the "tail-test") and compare the operating characteristics of the proposed approach with approaches currently used in practice. In contrast to the currently used approaches, we note that our proposed methodology is compatible with the proposed understanding of comparability and has, compared to other frequently applied range-based approaches, the advantage of being a formal statistical testing procedure which controls the patient's risk and has reasonable large-sample properties.

KEY WORDS: analytical studies; biosimilarity; equivalence testing; manufacturing change; quality attributes.

INTRODUCTION

Biologics are large-molecule drugs that have brought life-changing improvements to the health of patients in many important disease areas. A biosimilar (the test product) is developed and approved as a comparable version of an already marketed biologic (the originator or the reference product) and can be sold after the patent or other statutory exclusivity period of the originator has expired (1). For the showing of biosimilarity, regulators recommend a step-wise approach. The establishing of comparability at the quality/analytical level, i.e., demonstrating the similarity of the biosimilar and the originator at the structural and functional level, is generally seen as the first and most important step for the approval of biosimilars (2). The use of statistics in the comparability exercise of quality attributes is currently in the focus of attention because both the EMA and the FDA are working on regulatory guidances on this topic. However, there is still an ongoing controversial discussion on the question of if and how statistical approaches can be used for supporting the comparability claim. This is evident by the fact that the FDA has withdrawn its published draft guidance on this topic after reviewing the received comments (3).

We note that analytical studies are also the main piece of evidence for demonstrating comparable product quality after a manufacturing change. The main difference between the two types of assessment lies in the amount of available data and in the magnitude of the change. We focus on biosimilars in this paper, but our conclusions may also be relevant for manufacturing changes.

In this paper, we discuss several quantitative methodologies for performing a comparability assessment from a statistical perspective. This paper builds on the work of (4,5) who proposed and studied statistical methodologies for analytical similarity assessment using null and alternative hypotheses based on equivalence of the mean value. However, in this paper, we first take a step back and critically discuss which situations should be considered as "comparable" and which situations should be considered as "not comparable." This is motivated by observations made during biosimilar development in practice (6–8) and by the

Electronic supplementary material The online version of this article (https://doi.org/10.1208/s12248-018-0275-9) contains supplementary material, which is available to authorized users.

[1] Novartis Pharma AG, 4056, Basel, Switzerland.
[2] Sandoz GmbH, Kundl, Austria.
[3] Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.
[4] To whom correspondence should be addressed. (e–mail: byron.jones@novartis.com)

relevant ICH guidelines which request that the quality attributes are controlled to be within limits (one-sided, e.g., for impurities) or ranges (two-sided) and not in terms of mean values. A first proposal along those lines was made by (9,10), but their proposed approach contained also a mean value constraint and was therefore not fully focused on the range. In addition, it was not introduced as a formal inferential testing procedure. The aims of this paper are twofold: on the one hand, we intend to illustrate the inappropriateness of currently available statistical methodology for quality assessments if our understanding of comparability is used. On the other hand, we aim to present a more suitable alternative approach: we introduce the so-called tail-test which is an inferential methodology for comparability claims which is compatible with the ICH guidelines in the sense that a range-type hypothesis is tested. The operating characteristics of the tail-test are compared to other available approaches.

MATERIAL AND METHODS

In this section, we first introduce our understanding of comparable and not comparable settings. Afterwards, we present our proposal for an inferential statistical approach (the so-called tail-test) for quality assessment which is in line with the aforementioned definition of comparability. Next, we list the selected statistical methodologies which are most often used in practice for establishing comparability. These approaches are compared to the tail-test in a simulation study whose settings are described in the last part of this section.

Let X_R and X_T be random variables that represent the measurement of the quality attribute of Reference and Test, respectively. In this paper, we assume that both sets of random variables follow a normal distribution with mean μ_R and standard deviation σ_R for Reference and μ_T and standard deviation σ_T for Test. All observations are assumed to be independent and identically distributed. It should be emphasized that these assumptions might not be valid in practice for all quality attributes. However, these are the typical assumptions for many statistical approaches and that is why we have decided to use this simplified scenario. Let $x_1,\ldots, x_n$ be the realizations of X_T and $y_1,\ldots, y_m$ be the realizations of X_R.

A Meaningful Characteristic of Interest for the Comparability Assessment of Quality Attributes

When two populations are compared, the focus is typically on comparing the locations of the two populations, which are given, for example, by the mean values. Then, two populations would be considered comparable if the difference in their expected mean values is "small." This is also what is proposed explicitly by Tsong *et al.* (4) and in the withdrawn FDA's draft guidance (3) on statistical approaches for comparability in biosimilar development for the most critical quality attributes. In the EMA draft reflection paper on comparability assessment of quality attributes (11), the mean value is also a frequently used example for the characteristic of interest. However, in routine manufacturing, the mean value of observations for quality attributes across lots of the test or reference material is of no relevance for defining acceptable quality. The ICH Q8 guideline (12), for example, requests that "a CQA [critical quality attribute] is a physical, chemical, biological, or microbiological property or characteristic that should be within an appropriate limit, range, or distribution [e.g., a particle size distribution for solids] to ensure the desired product quality."

The shift from the mean value as the characteristic of interest to a range-type point of view is also motivated from experience in practice in biosimilar development: since comparability according to the ICH guidelines is currently only defined in terms of limits or ranges, it is not guaranteed that the mean value of the reference product is constant over time. Therefore, even the reference product might not fulfill a comparability criterion based on the mean value if different manufacturing time frames are compared. Using the mean value as the characteristic of interest for the comparability exercise of the biosimilar can be seen as an uncontrollable risk for the developer of the biosimilar and we illustrate this in Fig. 1: a biosimilar developer might develop its production process so that the mean values are comparable to the reference product lots which are available at the start of the biosimilar development (mean values of Test and Reference are identical in the beginning in Fig. 1, left hand panel). Afterwards, the reference product may undergo manufacturing changes and, while still being within specification limits (13) that were agreed with the regulatory authorities, the mean value might shift downwards (starting from time $t = 21$ in Fig. 1). If the biosimilar developer compares the mean value of the proposed biosimilar to the mean value of the reference product at a point in time after the manufacturing change, the mean values might not agree and lead to the false conclusion that the biosimilar and the reference product are not comparable—even though the biosimilar was comparable to the reference product at the time the development of the biosimilar started.

This is indicated in the left panel of Fig. 1 by the solid lines which give the cumulative mean values for Test and Reference: while both lines are at a comparable level at the beginning (e.g., at $t = 10$), the cumulative mean value of Reference is shifting downwards indicating a clear separation of the mean values at time $t = 40$. Therefore, a test on equivalence of mean values for all observations until $t = 40$ would, most likely and depending on the predefined equivalence margin, not allow a claim of comparability. On the other hand, if we focus on the distributions themselves (right hand panel in Fig. 1), it is clear that the range of Reference completely covers the range of Test and therefore the two products would be considered equivalent if a range-type criterion would be applied. It is important to note that shifts in the mean value are not only of theoretical interest, but this pattern has already been observed in practice and reported in publications (6–8).

We emphasize that comparing the empirical distribution functions instead of the mean values of Test and Reference, for example, with the Kolmogorov-Smirnov test (14), is not appropriate in this situation, since a smaller variability of Test than of Reference is acceptable (see, for example, ICH Q10, (15)). That is why, in this paper, we focus on a range-type hypothesis: we consider two products to be comparable if the tails of the distribution of Test are, at most, "slightly" heavier than the tails of Reference, whereupon slightly is more concretely defined via the equivalence margin c (see below).

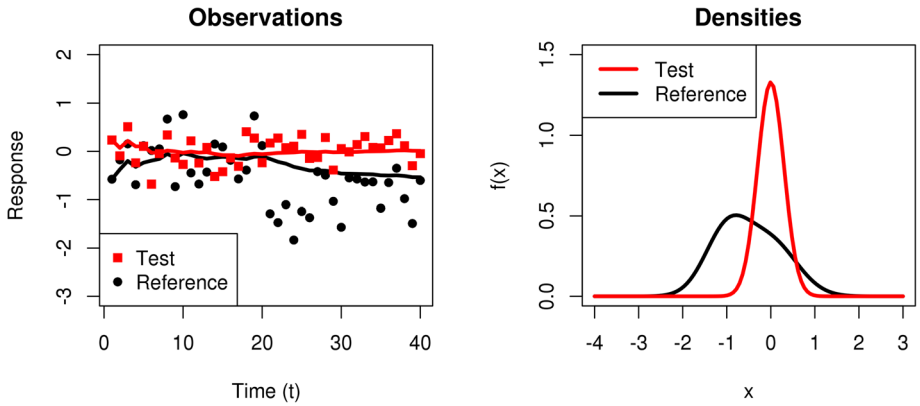

Fig. 1. Potential impact of manufacturing changes on a comparison of the mean value: the left hand panel shows (simulated) responses of a quality attribute displayed over time. The solid lines indicate the cumulative mean value (the mean value considering all observations until a specific point in time). The right hand panel shows the true, marginal densities that were used for generating the data

For that, let R_x be the x%-quantile of the Reference product, i.e., the value such that x% of the probability distribution of Reference lies below it. Then, we test the null hypothesis:

$$H_0 : 2q - \left(\left(P\left(X_T < R_q\right)\right) + P\left(X_T > R_{1-q}\right)\right) < c, \quad (1)$$

versus the alternative hypothesis:

$$H_1 : 2q - \left(\left(P\left(X_T < R_q\right)\right) + P\left(X_T > R_{1-q}\right)\right) \geq c,$$

where c is the equivalence margin (chosen, in discussion with subject matter experts, to be appropriate for the quality attribute used in the comparison) and q is the considered quantile with $q \in [0, 0.5]$ which has to prespecified. Reasonable choices are, for example, $q = 0.05$ or $q = 0.1$. Figure 2 provides an aid to understanding the logic behind these hypotheses.

In Fig. 2, the quantiles R_q and R_{1-q} are taken from the distribution of Reference (left hand panel, dotted lines). Then, the probability mass of the distribution of Test which is outside of the identified quantiles is calculated, i.e., the probability of observing a value of Test which is even more extreme than the estimated quantile of Reference (colored parts in the right hand panel, denoted by $C_l := P\left(X_T < R_q\right)$ for the red area and by $C_u := P\left(X_T > R_{1-q}\right)$ for the blue area). We consider the total probability to observe a value of Test outside of the quantiles of Reference, i.e., we focus on the sum of C_l and C_u and compare this to $2q$ (which is, by construction, the probability that a value of Reference is more extreme than the quantiles R_q and R_{1-q}, respectively).

If the tails of Reference are much heavier than the tails of Test, then the sum of the area of the colored parts is much smaller than $2q$. Thus, for claiming comparability, it is required that the difference between $2q$ and the sum of C_l and C_u is large. The value c gives the equivalence margin: one could set this value to $c = 0$ which means that we test if the sum of the probability mass below the q-quantile and above the $(1-q)$-quantile of Reference is larger for Reference than that for Test (i.e., "full" comparability). However, in the context of equivalence and non-inferiority testing, one generally allows for a margin in the size of a difference which reflects the amount that one would not consider two products to be different from a clinical point of view (16). For example, in the approach proposed by Tsong *et al.* (4), a difference in the mean value of $1.5\sigma_R$ is proposed as the equivalence margin. Therefore, in order to stay aligned with current practice, we allow that the tails of Test can be slightly heavier than the tails of Reference and still claim equivalence. The acceptable difference is given by the prespecified constant c which has to be chosen in discussion with subject matter experts.

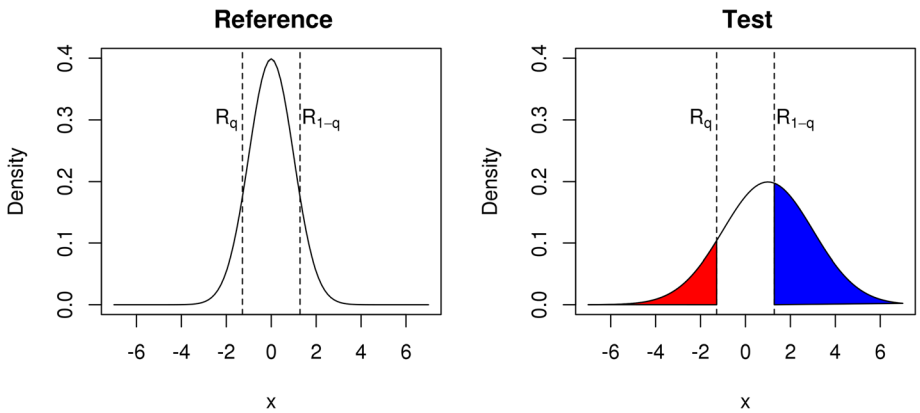

Fig. 2. The proposed null and alternative hypotheses. The left hand panel shows the density of Reference which is used for estimating the quantiles. The dashed lines give the estimated quantiles (R_q, R_{1-q}). The sum of the colored areas is compared to $2q$ (C_l, C_u)

The Tail-Test

The test statistic, which we introduce in this section, uses the condition stated in Eq. (1) as the null hypothesis and therefore considers the probability that a more extreme value for Test than the chosen q- and $(1-q)$-quantiles of Reference is observed. We refer to the proposed approach as the tail-test. The tail-test assumes normally distributed and independent data. However, the general idea of the approach can easily be adjusted to allow for other types of distribution, e.g., for a mixture distribution in which potential manufacturing changes are incorporated as long as sufficient data are provided for allowing a reliable estimation of the parameters of the distribution. It may, in principle, also be possible to take into account information on the lot structure of the data which can be useful if substantial lot-to-lot variability is expected. This is also discussed by the FDA (3). The tail-test uses the following procedure:

1. Estimate separately the mean value and variance of Test T and Reference R, assuming normal distributions $(\hat{\mu}_R, \hat{\mu}_T, \hat{\sigma}_R^2, \hat{\sigma}_T^2)$.
2. Calculate the q- and $(1-q)$-quantiles of Reference using the fitted normal distribution from step 1, i.e., the quantiles are given by:

$$\hat{R}_q = \hat{\mu}_R + z_q\,\hat{\sigma}_R \text{ and } \hat{R}_{1-q} = \hat{\mu}_R + z_{1-q}\,\hat{\sigma}_R,$$

where z_q is the q-quantile of the standard normal distribution, e.g., for $q = 0.05$, $z_{1-q} = 1.64$.

3. Calculate the probability that, assuming the fitted normal distribution of T, an observation of T is observed that is smaller than the lower quantile $\hat{R}_q$ or larger than the upper quantile $\hat{R}_{1-q}$. For that, let $X_{T,e}$ be a normally distributed random variable with mean value $\hat{\mu}_T$ and standard deviation $\hat{\sigma}_T$. Then, these probabilities are:

$$\hat{C}_l := P\left(X_{T,e} \le \hat{R}_q\right) \text{ and } \hat{C}_u := P\left(X_{T,e} \ge \hat{R}_{1-q}\right).$$

4. The difference between $2q$ and the sum of $\hat{C}_l$ and $\hat{C}_u$ serves as the test statistic w which is a realization of a random variable W, i.e.,

$$w = 2q - \left(\hat{C}_l + \hat{C}_u\right).$$

The realization of the test statistic is compared to the $(1-\alpha)$-quantile of the distribution of W under the null hypothesis. If the observed value w is larger than this critical value, comparability can be claimed. The derivation of the theoretical distribution under the null hypothesis is mathematically challenging and we therefore use critical values $W(\alpha, c, q, m, n)$ which are obtained using simulation. A Shiny-App (17) is available at https://nballarini.shinyapps.io/tailTest/ which gives a graphical interface for calculating the test decision with the tail-test for real datasets. For readers who prefer an implementation in R (18), we also provide code as supplementary material on the publisher's webpage.

Numerical example: We set $q = 0.1$, choose a significance level of $\alpha = 0.05$, and an equivalence margin of $c = -0.1$. We then assume that $n = 20$ measurements of Test and $m = 20$ measurements of Reference were taken and the estimated mean values and standard deviations (Step 1) are:

$$\hat{\mu}_R = -0.18, \hat{\mu}_T = -0.22, \hat{\sigma}_R^2 = 0.81, \hat{\sigma}_T^2 = 0.27$$

In Step 2, the quantiles of the Reference product are calculated where the q-quantile of the standard normal distribution for $q = 0.1$ is given by $z_q = -1.28$ and the $(1-q)$-quantile is given by $z_{1-q} = 1.28$:

$$\hat{R}_{0.1} = \hat{\mu}_R + z_q\,\hat{\sigma}_R = -0.18 - 1.28 \cdot 0.9 = -1.33$$
$$\hat{R}_{0.9} = \hat{\mu}_R + z_{1-q}\,\hat{\sigma}_R = -0.18 + 1.28 \cdot 0.9 = 0.97.$$

Then, in Step 3, the probability to observe a value of Test which is larger or smaller, respectively, than these quantiles is calculated. This is done by using the distribution function of a normally distributed random variable with the estimated parameters $(\hat{\mu}_T = -0.22,\ \hat{\sigma}_T^2 = 0.27)$ and comparing it to the estimated quantiles of the Reference product:

$$\hat{C}_l = P(X_{T,e} \le \hat{R}_q) = P(X_{T,e} \le -1.33) = 0.02$$
$$\hat{C}_u = P\left(X_{T,e} \ge \hat{R}_{1-q}\right) = 1 - P(X_{T,e} \le 0.97) = 0.01.$$

We also calculate the realization of the test statistic W:

$$w = 2q - \left(\hat{C}_l + \hat{C}_u\right) = 0.17.$$

The critical value is simulated with the provided Shiny-App or R-code and we obtain $W(0.05, -0.1, 0.1, 20, 20) = 0.08$. We note that $W(0.05, -0.1, 0.1, 20, 20) < w$ and therefore claim comparability.

Approaches Frequently Used in Practice for Comparability Claims

In the following, we briefly introduce the statistical approaches which are currently used in practice. We refer to (19) for a comprehensive overview of statistical approaches in quality assessment. The approaches described in the following will be compared to the tail-test in a simulation study.

Tiered Approach (Tsong et al. (4))

Tsong *et al.* (4) advocate a tiered approach for the assessment of quality attributes: the quality attributes which are considered to be most critical are called Tier 1 attributes and should be assessed using a formal equivalence test for the mean value. Quality attributes which are less important are grouped as Tier 2 and are compared with so-called quality ranges and the least relevant quality attributes (Tier 3) are analyzed descriptively. The way the assessments are grouped into the tiers is a topic of debate and will not be considered in this paper; we refer to (20) for more on this. In this paper, we will only focus on Tier 1 and Tier 2 attributes. The statistical

approaches discussed in this section are comparable to the FDA's approach in the withdrawn draft guidance (3). The Tier 1-tests aims for a decision on:

$$H_0 : |\mu_R - \mu_T| \geq \delta,$$

against the alternative hypothesis:

$$H_1 : |\mu_R - \mu_T| < \delta.$$

The equivalence margin δ is proposed to be $\delta = 1.5\sigma_R$. For a level α-test, it is suggested that a $(1-2\alpha)$ confidence interval is calculated and equivalence is claimed if the confidence interval fully lies within the interval:

$$\left[-1.5\sigma_R, 1.5\sigma_R\right].$$

In practice, the true standard deviation σ_R is not known and is therefore replaced by its estimated value $\hat{\sigma}_R$. We refer to this test as Tier 1-test (T1-test) in the rest of the paper.

Quality ranges should be used for Tier 2 attributes (4). For that, first, a quality range is calculated:

$$\left[\hat{\mu}_R - X\hat{\sigma}_R, \hat{\mu}_R + X\hat{\sigma}_R\right],$$

where the factor X needs to be justified and discussed with the regulatory authority. Commonly, a value between 2 and 5 is used and some discussion on the choice of X is provided in (21). For claiming comparability, it is required that "a large proportion, say 90%, of the test results from the biosimilar lots [fall] within an acceptance range determined by the reference product" (4). We refer to this approach as the quality range-approach (qr-approach) in this paper.

Min/Max-Approach

The min/max-approach is motivated by the idea that as all batches of Reference are already on the market, and therefore are obviously considered safe and efficacious, a batch of Test can be considered acceptable as long as it lies within the range of Reference. Therefore, only the most extreme values have to be compared. More formally, we define:

$$\hat{t}_{\min} = \min_{i=1,\ldots,n} x_i; \quad \hat{t}_{\max} = \max_{i=1,\ldots,n} x_i$$

and

$$\hat{r}_{\min} = \min_{i=1,\ldots,m} y_i; \quad \hat{r}_{\max} = \max_{i=1,\ldots,m} y_i$$

If $\hat{t}_{\min}$ is larger than $\hat{r}_{\min}$ and $\hat{t}_{\max}$ is smaller than $\hat{r}_{\max}$, then comparability has been established with the min/max-approach (mm-approach).

Prediction Interval in Tolerance Interval

A combination of prediction and tolerance intervals was introduced by Boulanger (22) who proposed basing the decision on comparability on the question as to whether future batches of Test will lie within the range of Reference with a prespecified level of confidence. Therefore, this approach goes in the direction of a range-type hypothesis even though no explicit hypotheses are stated. Boulanger (22) proposed defining the acceptable interval (the equivalence margin) using a β-content γ-confidence tolerance interval (23) for Reference.

For making a decision if Test and Reference are comparable, a β-prediction interval is estimated (24) and compared to the calculated tolerance interval. We will refer to this approach as the prediction-interval-in-tolerance-interval-approach (piti-approach). We note that (9) also proposed an approach based on tolerance intervals. However, their approach assumes a dependency structure between observations from the same lot which is not compatible with our assumption of independent and identically distributed random variables.

Setup of the Simulation Study

In the following section, we illustrate the operating characteristics of the tail-test compared to the other introduced statistical approaches using our understanding of comparability, i.e., we will simulate datasets under the above-defined null (see Eq. (1)) and alternative hypotheses and compare the rejection rates. We consider sample sizes relevant for biosimilar development. For that, we assume an equal sample size of

$$n = m = 5, 10, 25, 100$$

observations for both Test and Reference. The motivation for the inclusion of one sample size which is unrealistically high is to allow a check on the asymptotic (large sample size) properties of the statistical methodologies: a reasonable statistical approach should reward a large sample size and have the property that if the sample size increases, this should make it more likely that the correct test decision is made.

We discuss four different scenarios for these sample sizes assuming independent and identically normally distributed observations. The analyzed scenarios are comparable to some of the scenarios analyzed by (22). A graphical illustration of the scenarios is given in Fig. 3.

1. Distributions of Test and Reference are identical (correct decision: comparable; $\mu_R = \mu_T = 0$ and $\sigma_T = \sigma_R = 1$).
2. The mean values of Test and Reference are different and the variability of Test is much smaller than that of Reference so that the distribution of Reference completely covers the distribution of Test (correct decision: comparable; $\mu_R = 0$, $\mu_T = 0.9$, $\sigma_R = 1$ and $\sigma_T = 0.4$).
3. The mean values of Test and Reference are identical, but the variability of Reference is smaller than the variability of Test (correct decision: not comparable; $\mu_R = \mu_T = 0$ and $\sigma_R = 1$ and $\sigma_T = 2.15$).
4. The variances of Test and Reference are identical, but the distribution of Test is shifted so that the situation is under Tsong's null hypothesis (4) (correct decision: not comparable; $\mu_R = 0$ and $\mu_T = 1.5$ and $\sigma_R = \sigma_T = 1$).

Note that the judgements of the "correct decision" were made using our understanding of comparability with $q = 0.05$

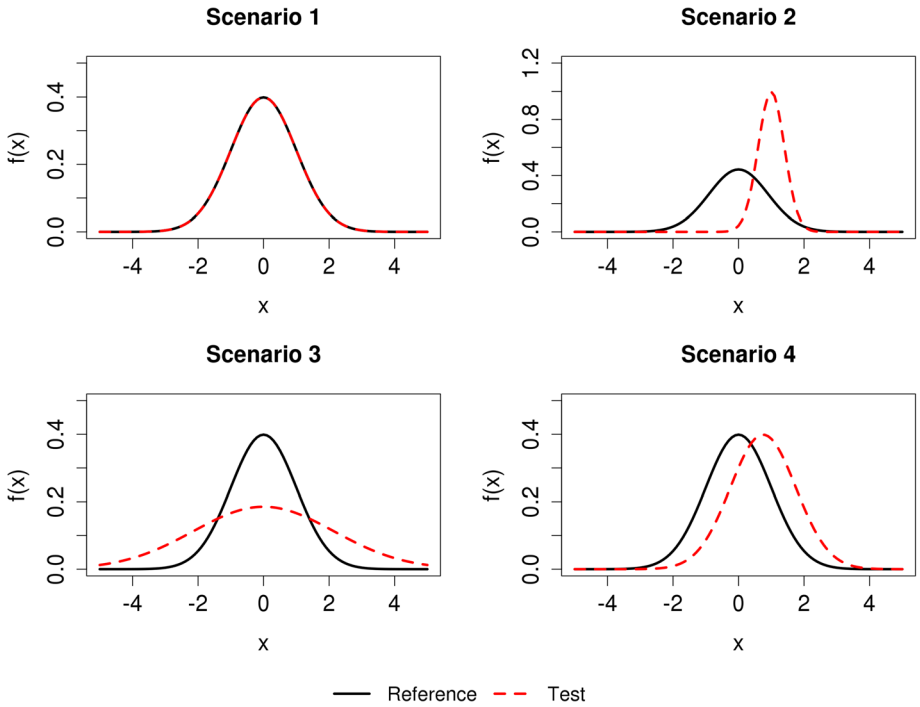

Fig. 3. The densities of Test and Reference for the four scenarios. Note: the *y*-axes do not show the same ranges

and an equivalence margin of $c = -0.34$, which corresponds to the value of the test statistic W under Tsong's null hypothesis (4) ($\mu_R = 0$, $\mu_T = 1.5$, $\sigma_R = \sigma_T = 1$). This setting is also used for deriving the critical value for the tail-test in order to ensure comparability with the T1-test.

There exists selectable parameters (tuning parameters) in the qr-approach (X) and in the piti-approach (β, γ) which allows us to calibrate the approaches for a specific value of n and m such that the Type I error rate in Scenario 4 is controlled at $\alpha = 0.05$. Even though, if applied in practice, one would use subject matter expertise or simulation studies for choosing these tuning parameters, making the Type I error rates comparable to the ones of the T1-test and the tail-test allows a meaningful comparison of the power profiles.

The mm-approach does not require tuning parameters by nature. However, we note that the Type I error rate in Scenario 4 is roughly controlled by the nature of the methodology for $m = n$. More details on the calibration of the approaches can be found in the online supplement.

For all four described scenarios, 1000 datasets for Test and Reference were simulated and the numbers of claims of comparability for all introduced statistical methods were recorded. All simulations were performed with R 3.2.3 (18). For the calculation of the tolerance intervals, the *R*-package tolerance (25) was used.

RESULTS

In Fig. 4, we show the percentages of times in which comparability is claimed dependent on the scenario for the four selected sample sizes. Scenario 4 is omitted since the comparability of the rejection rates under Scenario 4 was enforced with the calibration (see above). From the remaining scenarios, Scenarios 1 and 2 are the situations in which a claim of comparability is desirable whereas Scenario 3 is a setting in which the products are not truly comparable (low rate of comparability claims desirable).

We first consider the T1-test and note that the rate of comparability claims is among the highest in Scenario 1 and is reasonably high also for small sample sizes. For example, for $N = 10$, the rate of comparability claims is already higher than 80%. On the other hand, the rate of comparability claims is very low in Scenario 2. This is expected since the focus of the T1-test is on the mean value, but it is still a highly undesirable feature if a range type hypothesis is considered. Even more importantly, the rate of comparability claims is high in Scenario 3 and the rate of comparability claims in this situation is increasing with an increasing sample size and reaches 100% for $N = 100$. This illustrates that the T1-test is not an appropriate approach if comparability is interpreted using a range-type hypothesis.

Next, we consider the mm-approach. This approach shows high rejection rates under equal sample sizes if the variability of Test is much smaller than that of Reference (Scenario 2). Also, the patient's risk (false positive rate, Scenario 3) seems to be well-controlled. However, if the distributions of Test and Reference are identical, the rate of comparability claims is low (Scenario 1). This can be easily explained by calculating the expected rejection rate: assuming that the maximum and minimum would be independent, there is a 50% chance that the maximum of Test is smaller than the maximum of Reference and a 50% chance that the minimum of Test is larger than the minimum of Reference leading to an expected rejection rate of 25%. Since in fact the maximum and minimum are correlated, a slightly lower rejection rate is expected and also observed in the presented simulation study. In total, the mm-approach can only be recommended if the company is highly confident that the variability of Test is much smaller than the variability of Reference.

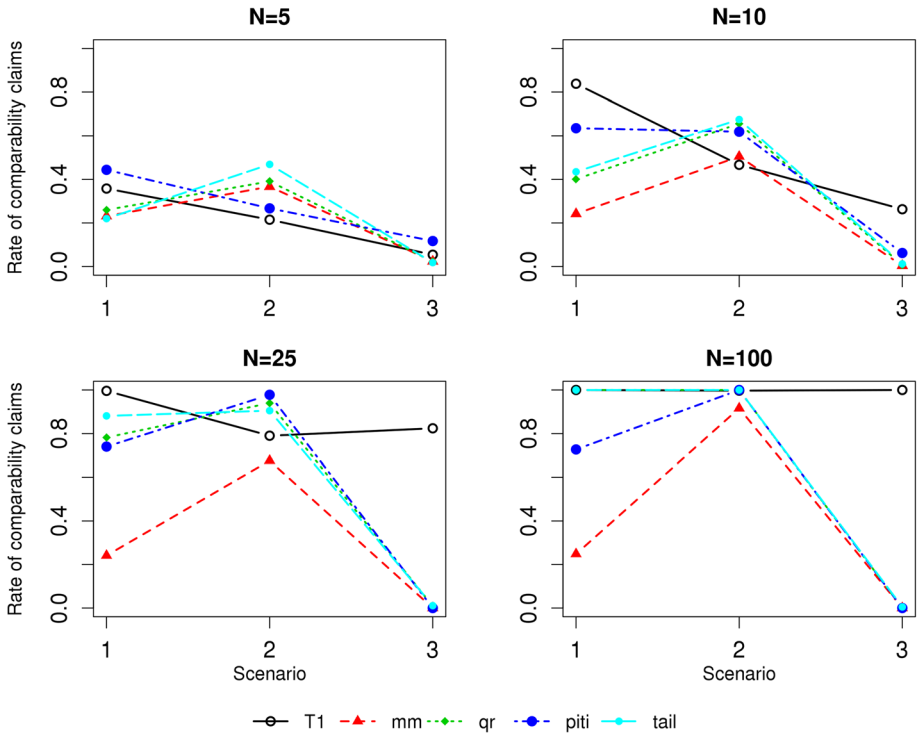

Fig. 4. Rate of comparability claims dependent on the number of observations of Reference and Test ($N = m = n$) for the tail-test (denoted as tail) compared to the four standard methodologies (T1, Tier 1-test; mm, min/max-approach; qr, quality range-approach; piti, prediction-interval-in tolerance-interval-approach). The scenarios refer to the settings described above (Scenarios 1 and 2, comparable; Scenarios 3, not comparable)

The performance of the piti-approach highly depends on the sample size: for $N = 5$, the rate of comparability claims is more than 11% in Scenario 3 (not comparable is the correct decision). Therefore, the rate of false positive decisions is too high for so small sample sizes and the piti-approach cannot be recommended for $N = 5$. For $N = 10$, the rate of comparability claims is already high in Scenario 1 and in Scenario 2 if compared with other approaches. However, the rate of comparability claims is still only around 60% which indicates that this is not a sufficient sample size in practice. The rate of comparability claims does not reach 80% even for $N = 100$ in Scenario 1 indicating that increasing the sample size does not greatly influence the power after a certain threshold is reached. Therefore, this approach does not show reasonable asymptotic statistical properties. It can only be recommended if only a very limited sample size ($n = m \approx 10$) is possible due to practical reasons and one is willing to accept a low chance of success for that sample size.

The qr-approach behaves similarly to the mm-approach for $N = 5$. With an increasing sample size, the rate of comparability claims increases which is a highly desirable feature. Nonetheless, the rate of comparability claims is in none of the discussed scenarios the highest of the compared approaches. The rate of comparability claims in Scenario 3 is low as is desirable.

For the tail-test, we note good control of the rate of false-positive decisions in Scenarios 3. In Scenarios 1 and 2, we observe the desirable properties of an inferential methodology: the probability of claiming comparability increases with an increasing sample size and reaches 100% if $N = 100$ observations are included. In addition, the rate of comparability claims is among the highest for all sample sizes in Scenario 2. Compared to the mm-approach, a higher rate of comparability rates is also shown in Scenario 1 independent of the sample size. The rate of comparability claims is also higher or similar for Scenario 1 compared to the qr-approach. Here, it should specifically be noted that for medium sample sizes (e.g., $N = 25$), the rate of comparability claims is relevantly higher for the tail-test than for the qr-approach (89% *vs*. 76%). Compared to the piti-approach, the rate of comparability claims is lower for $N = 10$, but higher for $N = 25$. It is emphasized that the rate of comparability claims for $N = 10$ is for all approaches so low that it is not a sufficient sample size in practice. Therefore, in all situations in which the success rate is reasonable, the tail-test outperforms all comparators.

For this simulation study, we calibrated the qr-approach, the piti-approach, and the tail-test by the use of tuning parameters or critical values (see the online supplement for details) to ensure a similar rate of comparability claims in Scenario 4 for the different sample sizes. This is a step which improved the comparability between the approaches. While for the tail-test, it is obvious that the sample size needs to be taken into account in the calculation of the critical values, this is not typically done for the two other approaches when applied in practice. It is therefore important to emphasize that the sample size has a major impact on the rate of comparability claims in Scenario 4. In Fig. 5, we display the relative change in percent of the tuning parameter or critical values which led to similar rates of comparability claims (the parameter X for the qr-approach, the parameter β for the piti-approach, and the critical value for the tail-test) compared to the value of this parameter for $N = 100$. For example, the value of X for $N = 5$ which leads to a rate of comparability claims of 5% in Scenario 4 needs to be nearly twice the value

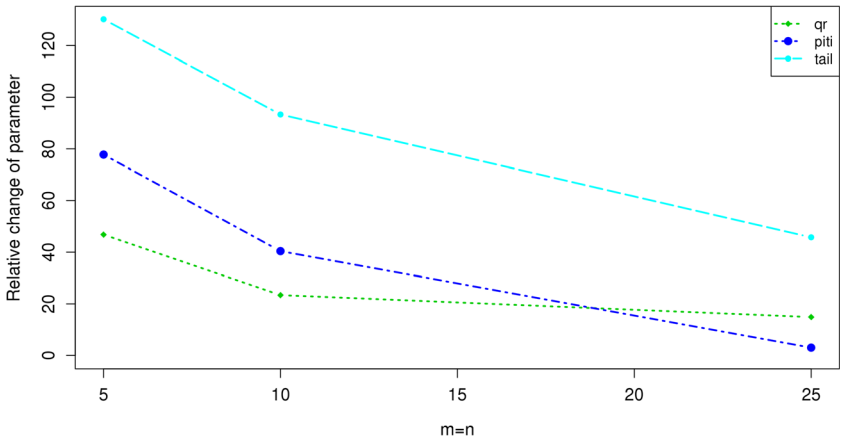

Fig. 5. Relative change of the tuning parameters/critical values for several sample sizes (compared to the tuning parameter/critical value for $n = m = 100$) for the qr-approach (denoted as qr), the tail-test (denoted as tail), and the piti-approach (denoted as piti)

which is required for $N = 100$. If one would simply use the same value of X for $N = 5$ that is used for $N = 100$, the rate of comparability claims increases from approximately 5% to more than 30%. Most importantly, the qr-approach and the piti-approach reject for a fixed tuning parameter more often if a smaller sample size is used which might discourage companies to use a reasonable sample size. This shows that also for an application in practice, the planned number of observations needs to be taken into account during the decision making process for the tuning parameters to ensure a low rate of comparability claims in scenarios which one would not consider to be comparable.

DISCUSSION

Aligning the statistical methodology with the aim of the analysis is crucial from a scientific point of view. Data analysis and statistical approaches can only give useful insights and support drug development if the scientific questions are correctly transferred into the statistical framework. For an improved alignment, we propose a set of statistical hypotheses in this paper which focus on comparing the ranges of Test and Reference instead of their mean values for establishing comparability. The tail-test is, to the best of our knowledge, the first proposal for an inferential statistical methodology for this class of hypotheses which is applicable in the area of comparability claims. The properties of the tail-test were illustrated in the simulation study in comparison to several approaches which are frequently used in practice for establishing comparability of quality attributes.

For the interpretation of the results of the simulation study, it is important to note that some of the compared methods are not inferential in nature: for example, one may interpret the min/max-approach as a descriptive approach since the test decision is purely based on the observed data only and the uncertainty of the estimation is not taken into account. That is why one can argue that it is in principle not possible to draw any conclusions regarding the underlying distribution. Nonetheless, when these approaches are applied in the context of a comparability assessment, a conclusion is made not only about the few observations which were studied, but also about the process itself, i.e., the question is answered if material produced with the analyzed manufacturing process (Test) will be sufficiently comparable to the established manufacturing process (Reference). That is why we assess the potential of all approaches, inferential or descriptive by nature, in answering questions related to the underlying population, i.e., we assume that the approaches aim to answer an inferential question.

In summary, our analysis showed that the proposed tail-test has desirable operating characteristics in a wide range of scenarios. In addition, we demonstrated that none of the other approaches can be recommended for the assessment of quality attributes without restrictions if a range-type hypothesis is used: the T1-test is not able to distinguish between comparable and not comparable settings. The qr-approach had lower or similar power than the tail-test in all situations. The rate of comparability claims for the piti-approach does not reach 100% even if $N = 100$ observations for Test and Reference are used and Test and Reference follow the same distribution. The piti-approach has therefore no reasonable large sample properties. The mm-approach has an acceptable control of the patient's risk. The developer's risk is only reasonably low if the variance of Test is much smaller than that of Reference. In other cases, the rates of a comparability claim are low—even if the distributions of Test and Reference are identical. Therefore, this approach is only applicable if the biosimilar developer is highly confident that the variability of Test will be lower than that of Reference, which can be a realistic scenario in biosimilar development (see, for example, (8)). Due to the high dependence of the test result on individual observations (no robustness), the application of the mm-approach requires a high level of prespecification of the analysis, e.g., the criteria for batch selection, the sample size, and the inclusion of all analyzed batches in the final analysis without the opportunity to select the most favorable results. It should also be noted that an approximately equal sample size of Test and Reference is required for the mm-approach for controlling the false positive rate.

It should be emphasized that there are also situations in which the tail-test is not recommended without modifications: these are the situations in which the assumption of independent and identical normally distributed observations is not justifiable. In these situations, a carefully conducted simulation study to assess the operating characteristics of potential approaches is required. It should, however, be emphasized that having representative data is relevant for all approaches,

not only when an inferential approach is applied, for which the representativeness of data is a formal requirement. Conclusions drawn with, for example, a descriptive approach based on only a few, but highly auto-correlated, measurements are also not a reasonable foundation for further steps in development. The sample should in any case be representative of the manufacturing process in the sense that the sample size is sufficient to capture the features of the underlying distributions. Other situations in which our proposal is not applicable are the ones with such small sample sizes such that the power of the proposed approach is too low. However, it should be emphasized that none of the other methodologies showed reasonable performance for small sample sizes over a wide range of scenarios.

CONCLUSION

The use of statistics in the comparability exercise of quality attributes is currently in the focus of attention because both the EMA and the FDA are working on regulatory guidances on this topic. In this paper, we proposed improving the alignment of the statistical hypothesis testing with scientific judgment by shifting away from the often-used comparison of mean values toward a range-based comparison. The tail-test, our proposal for an inferential methodology for this type of hypothesis, showed favorable performance over a wide range of scenarios in a simulation study. In addition, we demonstrated that statistical approaches which are currently used in practice all have serious weaknesses for the assessment of comparability if a range-based hypothesis is considered.

ACKNOWLEDGMENTS

We are grateful to Muhanned Saeed and Matej Horvat for fruitful discussions. We thank the three reviewers for providing well-thought-out comments which greatly improved this manuscript.

FUNDING INFORMATION

We acknowledge the funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 633567 and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 999754557.

COMPLIANCE WITH ETHICAL STANDARDS

Disclaimer The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government.

REFERENCES

1. Dranitsaris G, Amir E, Dorward K. Biosimilars of biological drug therapies. Drugs. 2011;71:1527–36.
2. FDA. Scientific considerations in demonstrating biosimilarity to a reference product. 2015. Available at http://www.fda.gov/downloads/DrugsGuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf. Accessed 03 Feb 2017.
3. FDA. FDA withdraws draft guidance for industry: statistical approaches to evaluate analytical similarity. 2018. Available at https://www.fda.gov/Drugs/DrugSafety/ucm611398.htm. Accessed 17 July 2018.
4. Tsong Y, Dong X, Shen M. Development of statistical methods for analytical similarity assessment. J Biopharm Stat. 2017;27:197–205.
5. Burdick RK, Thomas N, Cheng A. Statistical considerations in demonstrating CMC analytical similarity for a biosimilar product. Stat Biopharm Res. 2017;9:249–57.
6. Lamanna WC, Mayer RE, Rupprechter A, Fuchs M, Higel F, Fritsch C, *et al*. The structure-function relationship of disulfide bonds in etanercept. Sci Rep. 2017;7:3951.
7. Schiestl M, Stangler T, Torella C, Cepeljnik T, Toll H, Grau R. Acceptable changes in quality attributes of glycosylated biopharmaceuticals. Nat Biotechnol. 2011;29:310–2.
8. Kim S, Song J, Park S, Ham S, Paek K, Kang M, *et al*. Drifts in ADCC-related quality attributes of Herceptin®: impact on development of a trastuzumab biosimilar. MAbs. 2017;9:704–14.
9. Liao JJ, Darken PF. Comparability of critical quality attributes for establishing biosimilarity. Stat Med. 2013;32:462–9.
10. Liao JJZ. Statistical methods for comparability studies, Cham: Springer International Publishing; 2016. p. 675–94.
11. CHMP. Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development. 2017. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224995.pdf. Accessed 03 May 2018.
12. ICH. Pharmaceutical development Q8(R2). 2004. Available at https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q8_R1/Step4/Q8_R2_Guideline.pdf. Accessed 25 July 2018.
13. ICH. Validation of analytical procedures: text and methodology. 2015. Available at https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf. Accessed 03 May 2018.
14. Massey F Jr. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46:68–78.
15. ICH. Pharmaceutical quality systems. 2009. Available at https://www.fda.gov/downloads/drugs/guidances/ucm073517.pdf. Accessed 07 Sep 2018.
16. Ng TH. Noninferiority testing in clinical trials: issues and challenges. Boca Raton: CRC Press; 2014.
17. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: Web application framework for R. 2017. R package version 10.3.
18. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.
19. Liao JJZ. Comparability studies: statistical considerations. CRC Press; 2018. p. 1–17.
20. Vandekerckhove K, Seidl A, Gutka H, Kumar M, Gratzl G, Keire D, *et al*. Rational selection, criticality assessment, and tiering of quality attributes and test methods for analytical similarity evaluation of biosimilars. AAPS J. 2018;20:68.
21. 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;19:4–14.
22. Boulanger B. Assessment of analytical biosimilarity: the objective, the challenge, and the opportunities. 2017. Talk at ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop. Slides available at https://www.efspi.org/Documents/Events/Regulatory%20Meetings/2016/5.2.Bruno%20Boulanger_EFSPI_talk_analytical_biosimilarity_13SEP2016_V3.pdf. Accessed 03 May 2018.
23. Proschan F. Confidence and tolerance intervals for the normal distribution. J Am Stat Assoc. 1953;48:550–64.
24. Surhone L, Timpledon M, Marseken S. Prediction interval. Betascript Publishing; 2010.
25. Young DS. Tolerance: an R package for estimating tolerance intervals. J Stat Softw. 2010;36:1–39.