Stangler and Schiestl *AAPS Open* (2019) 5:4
https://doi.org/10.1186/s41120-019-0033-9

AAPS Open

RESEARCH Open Access

Similarity assessment of quality attributes of biological medicines: the calculation of operating characteristics to compare different statistical approaches

Thomas Stangler[1] and Martin Schiestl[2*]

Abstract

The comparison of quality attributes is a key element in the evaluation of both biosimilars and manufacturing process changes for biological medicines. Different statistical approaches are proposed to facilitate such evaluations. However, there is no regulatory consensus on a quantitative and scientifically justified definition and an underlying hypothesis of a statistically equivalent quality. The latter is essential to calculate operating characteristics of different approaches. This article proposes a hypothesis for establishing statistically equivalent quality which is concordant with current regulations. It also describes a tool which allows comparisons of different statistical approaches or tests by calculating the operating characteristics for false acceptance and false rejection rates of a claim for statistically equivalent quality. These error rates should be as low as possible to allow a meaningful application of a statistical approach in regulatory decision making. The described tool can be used to compare different statistical approaches for their suitability and may also facilitate the discussion and development of statistical approaches for comparing quality attributes in similarity assessments in general.

Keywords: Biosimilarity, Manufacturing changes, Quality attributes, Operating characteristics, Statistics

Introduction

Biosimilars are highly similar to a previously licensed biological reference product (U. S. Food and Drug Administration Guidance for Industry 2015; European Medicines Agency 2014). Equally, regulators require that biological medicines that undergo a manufacturing process change are demonstrated to be highly similar to the version of the medicine before the process change (ICH Harmonised tripartite guideline Q5E 2004). In both cases a thorough comparison of the quality attributes such as the physicochemical and functional attributes sets the foundation to establish the high similarity. Clinical studies are rarely needed for introduction of process manufacturing changes and are used in a tailored manner during biosimilar development, which puts the major burden of proof on the comparison of quality attributes (U. S. Food and Drug Administration Guidance for Industry 2015; European Medicines Agency 2014; McCamish and Woollett 2012). Notably, the definition of "high similarity" includes a certain discretion and allows for differences between the two products in quality attributes if sufficient product understanding is available to conclude that those differences are clinically meaningless. Many quality attributes may also show a certain but controlled variability between different production lots of a given product (Lamanna et al. 2017; Schiestl et al. 2011; Kim et al. 2017). Regulatory guidelines request that manufacturers control the critical quality attributes of biologics to stay within appropriate ranges or limits, so that the quality and clinical properties remain consistent over time (ICH harmonised tripartite guideline Q7 2000; ICH Harmonised tripartite guideline Q8(R2) 2009). Comparing quality attributes in a similarity exercise requires the comparison of different lots to describe and analyze the variability between them. Different non-inferential and inferential statistical approaches have been proposed with the aim to

* Correspondence: martin.schiestl@sandoz.com
[2]Sandoz GmbH, 6336 Langkampfen, Austria
Full list of author information is available at the end of the article

increase objectivity and robustness of these assessments (Tsong et al. 2017; Chow et al. 2016). The simplest approach is the visual display. Other approaches compare the test sample with observed (MinMax) or estimated ranges of the reference sample, such as xSigma or tolerance intervals. Yet another approach is equivalence testing of means. However, the application of statistics is not as easy as it may appear at the first glance. In 2018 FDA withdrew a dedicated draft guidance with specific proposals, which may illustrate the difficulties in establishing standards in this area (U. S. Food and drug administration 2018). A first caveat when applying statistical tests is the essential flexibility of the requirement of "high similarity" to allow for differences if they are clinically meaningless. Statistics may facilitate the detection of differences, e.g. in data distributions or ranges, but the determination whether or not these differences are clinically relevant is a scientific question that cannot be addressed by a statistical approach alone. This article focuses on the meaningful detection of statistical differences. In the event that differences are detected, the next step in the evaluation of claims for comparability of a manufacturing change or for biosimilarity, i.e. the determination of the clinical relevance of detected differences, requires the assessment of all relevant product and process knowledge, including structure-function relationships, understanding of the mode of action, safety data, and clinical experience with the product.

A comparison of the different statistical approaches or tests requires the calculation of the operating characteristics based on a clear hypothesis for accepting a claim for statistically equivalent quality. The concept of statistically equivalent quality is the scientific basis behind existing regulations for manufacturing process changes as well as the variability of quality attributes in routine production. This article proposes such a hypothesis and a tool which allows the calculation and comparison of the operating characteristics such as the average false acceptance rate and average false rejection rate. A false rejection means that a product is rejected although it fulfills our hypothesis for equivalent quality, whereas a false acceptance means that a product is accepted although it does not fulfill our hypothesis. The numbers for those error risks as calculated by the tool are relative and not absolute because they depend on the calculation parameters. However, they allow meaningful comparisons of the different statistical tests with regards to their utility in similarity exercises. For the purposes of simplicity within this article, we use the terms statistical approach and statistical test synonymously.

Materials and methods

Hypothesis for accepting a claim for statistical equivalence for the analyzed quality attribute

The variability of the reference product defines the acceptable quality for the test product population. Equivalence for the quality attribute is established if the population of the test product lies within the population of the reference product. The width of the population is described by 3σ because 3σ is commonly used as a threshold to describe a population. E.g. In statistical process control data points beyond 3σ are investigated because they might result from special cause variability and not belong to the population.

The population of the test product is therefore in the population of the reference product if $\mu_{test} - 3\sigma_{test} > \mu_{ref} - 3\sigma_{ref}$ and $\mu_{test} + 3\sigma_{test} < \mu_{ref} + 3\sigma_{ref}$. In other words, and assuming normal distributions, if at least the central 99.7% of the test product are within the central 99.7% of the reference product. The equivalence region described by this definition is illustrated as a triangle in Fig. 1.

The tool - calculation of the average false acceptance rates and average false rejection rate

Acceptance rates, i.e. likelihood of passing the test, are calculated by Monte Carlo methodology. Under the assumption of normally distributed data, the reference population and test population are distinguished by a relative difference in mean $(\mu_{test}\text{-}\mu_{ref})/\sigma_{ref}$ and the ratio of SD (standard deviation) $\sigma_{test}/\sigma_{ref}$. For any given sample size for the reference product and test product, n_{ref} and n_{test} respectively, n_{ref} and n_{test} samples are drawn repeatedly ($n_{sim} = 1000$) and randomly from the defined

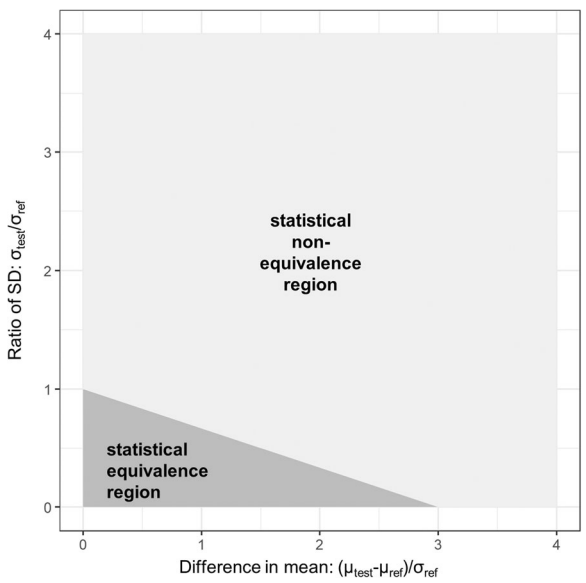

Fig. 1 Definition of the statistical equivalence region: The dark grey region is the defined true similarity region and indicates those combinations of SD ratio and difference in mean where the test population is within the reference population. The population width is described by μ ± 3σ, which translates into a triangle with the corners of SD ratio and difference in means of (0/0, 1/0, 0/3). The light grey region is the statistical non-equivalence (false similarity) region indicating where the test population is not considered to be within the reference population

reference and test population and evaluated using the statistical test (MinMax, 3Sigma, tolerance interval, equivalence test for means) for acceptance. The acceptance rate is calculated by the proportion of accepted samples to generated samples (n_{sim}). For any given sample size combination, acceptance rates can be calculated systematically for all relevant combinations of the difference in means and the ratio of SD. In this article, acceptance rates were calculated for a grid covering all combinations of the difference in means from 0 to 4 σ_{ref} with a stepsize of 0.1 σ_{ref} and ratio of SD from 0 to 4 with a step size of 0.1.

Calculated acceptance rates can be visualized by plotting the acceptance rate as a function of the difference in means $(\mu_{test}-\mu_{ref})/\sigma_{ref}$ and the ratio of SD $\sigma_{test}/\sigma_{ref}$ (see Additional file 1: Figure S1. for an example contour plot).

Average false acceptance rates (false positive) are calculated as an average of the acceptance rates for all grid points in the statistical non-equivalence region, which consequently represent false acceptance rates. Average false rejection rates (false negative) are calculated as an average of all rejection rates (1 − acceptance rates) for all grid points in the statistical equivalence region.

The code for these calculations is provided in the Additional file 1.

Statistical tests

a) MinMax: A MinMax range is defined by the lowest and highest value of a sample. The MinMax test is accepted if the MinMax range of the test sample is within the MinMax range of the reference sample ($min_{Test} > min_{Ref}$ and $max_{Test} < max_{Ref}$)
b) 3Sigma: the 3Sigma range is calculated for the reference sample as ($\mu_{ref}-3\sigma_{ref}$, $\mu_{ref}+3\sigma_{ref}$). The 3Sigma test is accepted if the MinMax range of the test sample is within the 3Sigma range.
c) Tolerance interval (TI): The tolerance interval is calculated for the reference sample as ($\mu-k^*\sigma_{ref}$,$\mu+k^*\sigma_{ref}$). The k-factor is calculated two-sided with a confidence level of 0.9 and a proportion of the population covered by the tolerance interval of $P = 0.99$. The tolerance interval test is accepted if the MinMax range of the test sample is within the tolerance interval calculated for reference sample.
d) Equivalence testing of means (EQT): A two one-sided t-tests' (TOST) procedure is used to test for equivalency of the means of the reference product and the test product. The equivalence margin is defined as $\delta = 1.5\ s_{ref}$ (standard deviation of the reference product sample), the Type I error probability is controlled at level $\alpha = 0.05$ for a conclusion of equivalence. The test is accepted if the $(1-2\alpha)100\% = 90\%$ confidence interval for the difference in the means is within $(-\delta, +\delta)$.

Results

The operating characteristics of a statistical test depend on the underlying hypothesis of statistical equivalence for the quality attribute, which, for this article, is fulfilled if the population of the test product lies within the population of the reference product. This hypothesis reflects the current regulation of manufacturing processes, which require that critical quality attributes are controlled within ranges or limits. Under the assumption of normally distributed populations, Fig. 1 illustrates the resulting region of statistical equivalence for a test product population which is distinguished from the reference population by a difference in means and the ratio of the distribution width (ratio of SD). The triangle illustrates the equivalence region as the area where the conditions for statistical equivalence as defined above are met. Average false rejection and average false acceptance rates were calculated for the different statistical tests as described in the Methods section and displayed in Fig. 2 which provides a comparison of the statistical tests for these error rates for sample size of n_{ref} and $n_{test} = 10$.

Figure 3 shows the impact of varying sample size for MinMax, 3Sigma and Equivalence testing of means on the average false acceptance and rejection rates. For

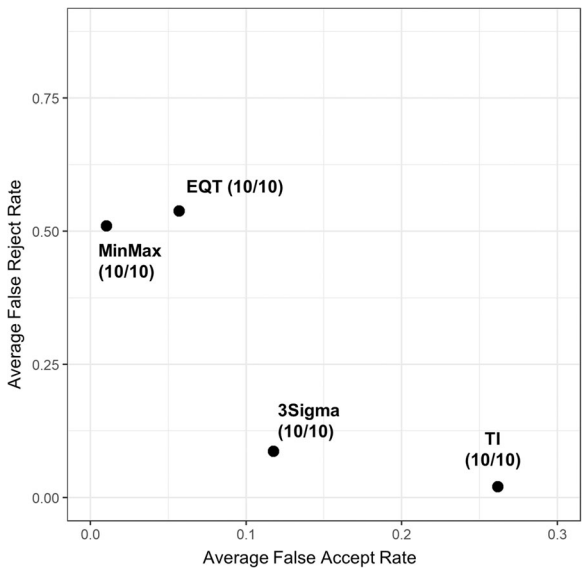

Fig. 2 Average operating characteristics: The average operating characteristics are shown for the tests MinMax, 3Sigma, equivalence testing of means and tolerance interval for a sample size of nref = 10 and ntest = 10. The horizontal axis represents the average false acceptance rate and represents the risk for a statistical false positive conclusion on similarity. The vertical axis represents the average false rejection rate and represents the risk for a statistical false negative conclusion on similarity

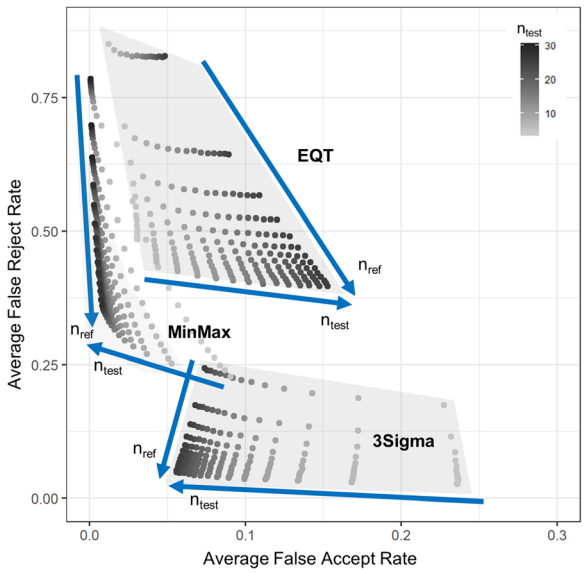

Fig. 3 Sample size dependency of average operating characteristics: The average operating characteristics are shown for different test and reference product sample sizes (n_{ref} = 4,6,8,...,30, n_{test} = 4,6,8,...30) for the tests MinMax, 3Sigma and equivalence testing of means. Arrows indicate the effect of increasing test product and reference product sample size (annotated as n_{test} and n_{ref}, respectively). The grey background polygons accentuate the area of operating characteristics for the individual tests

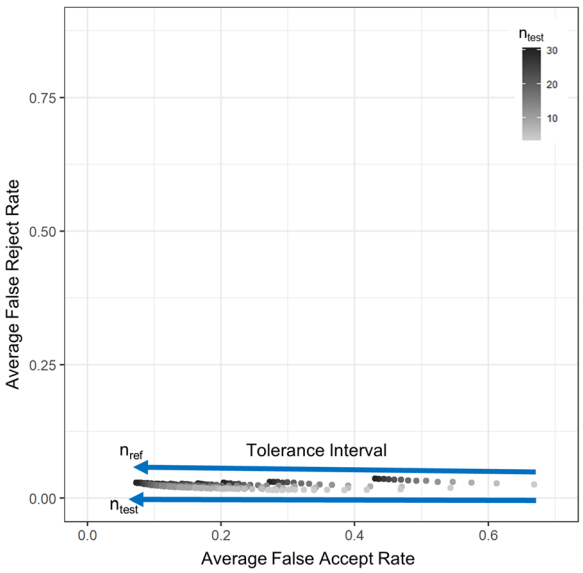

Fig. 4 Sample size dependency of average operating characteristics (Tolerance Interval): The average operating characteristics are shown for different test and reference product sample sizes (n_{ref} = 4,6,8,...,30, n_{test} = 4,6,8,...30) for the test Tolerance Interval. Arrows indicate the effect of increasing test product and reference product sample size (annotated as n_{test} and n_{ref}, respectively). In contrast to Fig. 3, the x-axis for the false acceptance rates is extended in order to visualize all data points

MinMax, increasing n_{ref} lowers the false rejection rate without a large impact on false acceptance rate. Increasing n_{test} reduces false acceptance rates but slightly increases false rejection rates. Similar trends with different magnitudes are observed for 3Sigma, where increasing n_{ref} reduces false rejection rates but also slightly reduces false acceptance rates. Increasing n_{test} reduces strongly false acceptance rates while it has only marginal impact on the false rejection rate. While increasing sample sizes reduce the false acceptance rate for MinMax and 3Sigma as expected, they increase the likelihood for passing the test, and associated false acceptance rates, for equivalence testing of means. This effect is especially pronounced with increasing test sample size. This different behavior of equivalence testing can be attributed to the lack of alignment of the equivalence test with the proposed equivalence hypothesis requiring that the population of the test product lie within the population of the reference product. Range-based tests are in general better aligned with this equivalence hypothesis. Tolerance interval testing shows generally a low false rejection rate but at small sample sizes there is also a high false acceptance rate. However, increasing sample sizes, especially for n_{ref} reduce the false acceptance rate to levels comparable with the other statistical tests (see Fig. 4).

Discussion

From a statistical viewpoint, without further knowledge of the impact of differences in quality attributes on safety and efficacy and without taking into account any risk mitigation by proper control strategy in manufacturing, the average false acceptance and rejection rates represent estimates for false positive and false negative decision of similarity between quality attributes of two products. Both error rates are important and should be as low as possible, however, a small false acceptance rate is even more desirable because it might impact risks posed to the patient, whereas a false rejection rate primarily impacts the risk for the manufacturer. The tool is therefore well suited to compare different statistical tests for its applicability in similarity assessments. Any specific application for a similarity exercise additionally requires consideration of potential multiplicity effects as typically many quality attributes are compared in parallel (Bretz et al. 2010). The tool also assumes normally distributed data and process variability without special cause variation, meaning that the analytical variability is negligible and the sample data do not shift over time. Non-normally distributed data and special cause variation require additional considerations with regard to sampling distributions and data evaluation.

The results provided in this article reveal that MinMax is a conservative approach with a low false acceptance rate, but it has a high false rejection rate. Equivalence testing has also a high false rejection rate and with increasing sample size a considerable false acceptance rate. The 3Sigma approach provides a more practical compromise of error rates, which further improves with larger sample size. Tolerance interval testing is only usable if sample size is sufficiently large.

A frequent practical question in the evaluation of similarity is on how many test samples are needed for robust decision making. The tool shows nicely that very small sample sizes can considerably increase the false acceptance rates for the range-based tests. The tool allows definition of acceptable sample sizes based on desired operating characteristics and/or to investigate alternative strategies to control the false acceptance rate.

For the equivalence test, on the other hand, an increasing sample size leads to greater precision in estimating the mean difference. In combination with the lack of alignment of the EQT with the equivalence hypothesis (test population in a reference population), this leads to an undesired increase of the false acceptance rate with increasing sample size.

While the examples illustrate the impact of sample size, the tool can also be used to assess the impact of other statistical testing parameters on the false acceptance and rejection rates. Finally, alternative hypotheses for statistical equivalence of the quality attributes can be easily assessed. For example, the equivalence area can be defined differently to allow a small difference in means when σ is the same on one side, but restrict the uncomfortable but also highly unlikely situation that a very narrow distributed test distribution is located in the far tail of the reference distribution. Such a hypothesis could define equivalence of the quality attribute if the central 95% if the test population are within the central 99% of the reference population. – see Additional file 1: Figure S2. For the operating characteristics of such an alternate hypothesis please see Additional file 1: Figure S3. (MinMax, 3Sigma, Equivalence testing of means) and Fig. 4 (TI).

Conclusion

Regulatory guidelines for biosimilar evaluation and comparability of process manufacturing changes require highly similar quality attributes between biosimilar candidate and reference medicine, and pre- and post-change product respectively (U. S. Food and Drug Administration Guidance for Industry 2015; European Medicines Agency 2014; ICH Harmonised tripartite guideline Q5E 2004). The definition of "high similarity" of quality attributes includes a range of variability. Even statistically significant differences could be acceptable if sufficient knowledge allows the conclusion that such differences are clinically meaningless. However, statistical approaches may facilitate the comparison of quality attributes by identification of statistical differences, which require further scientific evaluation before drawing a conclusion on whether a claim of high similarity is fulfilled or not. The tool presented in this article provides a means to calculate relevant operating characteristics such as error rates for average false acceptance (false positive) and false rejection (false negative) rates of different statistical approaches. Those properties allow a meaningful comparison and proper selection of statistical tests and may inspire research for novel statistical approaches for comparing quality attributes.

Additional file

Additional file 1: Figure S1. Contour plot overlay of acceptance rates for equivalence test (solid lines) and MinMax (dashed lines). Acceptance rates are plotted as a function of the difference in means (μtest-μref)/σref on the horizontal axis and the ratio of SD σtest/σref. on the vertical axis. Contour levels are equally spaced by 10% points and range from 10% (light grey) to 90% (dark grey). The figure was generated with nsim = 10,000. **Figure S2.** Definition of an alternative statistical equivalence region: The dark grey region is the true similarity region and indicates those combinations of ratio of SD and difference in mean where the central 95% of the test population are within the central 99% of the reference population. This results in a triangle with the corners of SD ratio and difference in means of (0/0, 1.32/0, 0/2.58). The light grey region is the statistical non-equivalence (false similarity) region. **Figure S3.** Sample size dependency of average operating characteristics for the alternative equivalence hypothesis: The average operating characteristics are shown for different test and reference product sample sizes (nref = 4,6,8,...,30, ntest = 4,6,8,...30) for the tests MinMax, 3Sigma and equivalence testing of means. The grey background polygons accentuate the area of operating characteristics for the individual tests. **Figure S4.** Sample size dependency of average operating characteristics for the alternative equivalence hypothesis (Tolerance Interval): The average operating characteristics are shown for different test and reference product sample sizes (nref = 4,6,8,...,30, ntest = 4,6,8,...30) for the test Tolerance Interval. In contrast to: **Figure S3.**, the x-axis for the false acceptance rates is extended in order to visualize all data points. R script for the calculation of average false acceptance and average false rejection rates. (DOCX 248 kb)

Abbreviations

EQT: Equivalence testing of means; FDA: Food and Drug Administration;; max_{Ref}: maximum value of reference product sample; max_{Test}: maximum value of test product sample; min_{Ref}: minimum value of reference product sample; min_{Test}: minimum value of test product sample; n_{ref}: number of reference product lots; n_{sim}: number of simulated and generated samples; n_{test}: number of test product lots; TI: tolerance interval; TOST: two one-sided t-test; α: significance level of Type I error; δ: equivalence margin; μ_{ref}: mean value of the reference product population; μ_{test}: mean value of the test product population; σ or SD: standard deviation; σ_{ref}: standard deviation of the reference product population; σ_{test}: standard deviation of the test product population

Acknowledgements

We would like to thank Matej Horvat, Franz Innerbichler from Novartis for their fruitful thoughts and discussions, and Hillel Cohen from Sandoz for his thorough review which greatly improved this manuscript.

Authors' contributions

TS described the hypothesis for establishing statistically equivalent quality. He invented the tool and made all calculations. MS provided the link to regulatory science and wrote the manuscript. Both authors developed the described examples. Both authors read and approved the final manuscript.

Funding
No funds were received to write this manuscript.

Availability of data and materials
All data and modes of calculation are part of the manuscript and the Additional files. The calculation method for the described tool is listed in the Additional files as complete R-code script (R Core Team 2018; Wickham 2017; Wickham 2016; Robinson 2016; Young 2010).

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors are employees of Sandoz and Novartis, which is developing, manufacturing and marketing biological medicines including biosimilars.

Author details
[1]Novartis, Sandoz GmbH, 6336 Langkampfen, Austria. [2]Sandoz GmbH, 6336 Langkampfen, Austria.

Received: 23 April 2019 Accepted: 29 August 2019
Published online: 10 September 2019

References

Bretz F, Hothorn T, Westfall P (2010) Multiple comparisons using R. 1. Chapman and Hall/CRC
Chow SC, Song F, Bai H (2016) Analytical similarity assessment in biosimilar studies. AAPS J 18:670–677
European Medicines Agency: Guideline on similar biological medicinal products CHMP/437/04 Rev 1. 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/10/WC500176768.pdf. Accessed 3 Apr 2019
ICH harmonised tripartite guideline: Good manufacturing practice guide for active pharmaceutical ingredients Q7. 2000. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q7/Step4/Q7_Guideline.pdf. Accessed 3 Apr 2019
ICH Harmonised tripartite guideline: Comparability of biotechnological/biological products subject to changes in their manufacturing process Q5E. 2004. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q5E/Step4/Q5E_Guideline.pdf. Accessed 3 Apr 2019
ICH Harmonised tripartite guideline: Pharmaceutical development Q8(R2). 2009. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q8_R1/Step4/Q8_R2_Guideline.pdf. Accessed 3 Apr 2019
Kim S, Song J, Park S, Ham S, Paek K, Kang M et al (2017) Drifts in ADCC-related quality attributes of Herceptin®: impact on development of a trastuzumab biosimilar. mAbs. 9:704–714
Lamanna WC, Mayer RE, Rupprechter A, Fuchs M, Higel F, Fritsch C et al (2017) The structure-function relationship of disulfide bonds in etanercept. Sci Rep 7:3951. https://doi.org/10.1038/s41598-017-04320-5
McCamish M, Woollett G (2012) The state of the art in the development of biosimilars. Clin Pharm Ther 91:405–417
R Core Team 2018. R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. URL https://www.R-project.org/. Accessed 25 July 2019
Robinson A. 2016. Equivalence: provides tests and graphics for assessing tests of equivalence. R package version 0.7.2. https://CRAN.R-project.org/package=equivalence. Accessed 25 July 2019
Schiestl M, Stangler T, Torella C, Čepeljnik T, Toll H, Grau R (2011) Acceptable changes in quality attributes of glycosylated biopharmaceuticals. Nat Biotechnol 29:310–312
Tsong Y, Dong X, Shen M (2017) Development of statistical methods for analytical similarity assessment. J Biopharm Stat 27:197–205
U. S. Food and drug administration: FDA withdraws draft guidance for industry: Statistical approaches to evaluate analytical similarity 2018. https://www.fda.gov/Drugs/DrugSafety/ucm611398.htm. Accessed 3 Apr 2019
U. S. Food and Drug Administration Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. 2015. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf. Accessed 3 Apr 2019
Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Publishing, New York
Wickham H. 2017. idyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse. Accessed 25 July 2019
Young DS. Tolerance: an R package for estimating tolerance intervals. J Stat Softw, 2010; 36(5): 1–39. http://www.jstatsoft.org/v36/i05/. Accessed 25 July 2019

Publisher's Note