Statistical principles in biosimilar development

Alison Balfour and **Susanne Schmitt**
Hexal AG, Holzkirchen, Germany

Correspondence to:

Alison Balfour
Associate Director
Hexal AG
Industriestr. 25
D-83607 Holzkirchen
Germany
alison.balfour@novartis.com

Abstract

Unlike new drug development where superiority over an active comparator or placebo often has to be proven, biosimilar development focuses on showing similarity of the proposed biosimilar to an already approved reference product. This affects the statistical aspects of clinical trials including choice of study design, endpoints, and analyses performed. In addition, there is a greater focus on margin justification and missing data imputation for efficacy. This article provides an overview of the statistical principles inherent to biosimilar development.

Lean clinical development programme

Biosimilar development is based on extensive physicochemical characterisation of the proposed biosimilar, followed by a lean clinical development programme to address any residual uncertainty about the similarity between the proposed biosimilar and the reference product. Typically, the clinical development programme is limited to two clinical studies: one pharmacokinetics/pharmacodynamics (PK/PD) similarity study and one confirmatory efficacy/safety/immunogenicity study. No dose finding study is usually conducted as the approved dose is known from the reference product.

Three arm active control PK/PD similarity study

The objective of the PK/PD similarity study is to demonstrate bioequivalence (show no clinically meaningful differences) in PK and/or PD between the proposed biosimilar and the authorised reference product. As different health authorities approve medicinal products in different regions, the authorised reference product may also vary by region, for example a US-licensed reference product versus an EU-authorised reference product. Due to this, the PK/PD similarity study usually includes three treatment arms: the proposed biosimilar, the EU-authorised reference product, and the US-licensed reference product. This results in three treatment comparisons: biosimilar vs EU reference, biosimilar vs US reference, and EU reference vs US reference (Figure 1).

Interval hypothesis testing

Statistically, PK similarity is demonstrated if the 90% confidence interval (CI) for the ratio of geometric means of test product to reference product for the PK parameter(s) – typically area under the curve from time zero to infinity (AUC_{inf}), maximum measured concentration (C_{max}), and area under the curve from time zero until the last quantifiable concentration (AUC_{last}) – falls entirely within the pre-defined margin of 0.80 to 1.25. This method is equivalent to conducting two 1-sided tests at the 5% level. If μT and μR respectively denote the population means for test and reference product for a particular endpoint, then the following null (H_0) and alternative (H_1) hypotheses are being tested:

$$H_0: {}^{\mu_T}/_{\mu_R} \leq 0.80 \; or \; {}^{\mu_T}/_{\mu_R} \geq 1.25$$

$$H_1: 0.80 < {}^{\mu_T}/_{\mu_R} < 1.25$$

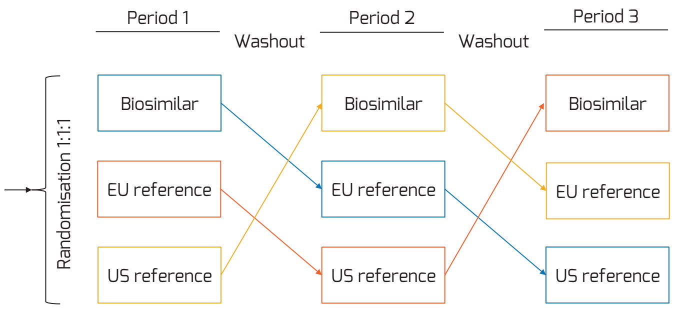


Figure 1. Typical PK/PD similarity crossover study design

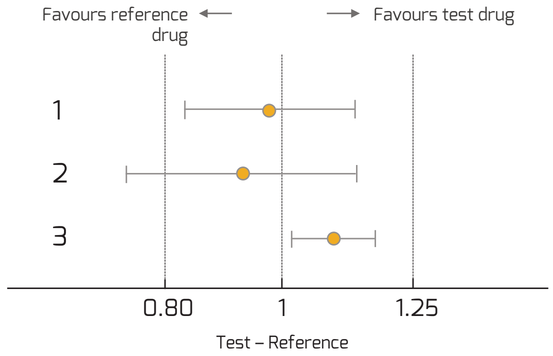


Figure 2. Examples of equivalence testing with confidence intervals.

1. Equivalence met: confidence interval contained entirely within margin of 0.80 to 1.25.
2. Equivalence not met: confidence interval partially outside the margin of 0.80 to 1.25.
3. Equivalence met, but additional explanation needed for why the confidence interval does not contain the equality point of 1.

variance (ANOVA) is performed on the log-transformed PK parameter and estimates for each treatment comparison are computed. For crossover studies the ANOVA model includes treatment sequence, treatment group, and period as fixed effects, and subject nested within treatment sequence as a random effect. For parallel group studies the ANOVA model should only include treatment group as a fixed effect. In addition, stratification factors used during randomisation and other important baseline characteristics may be used as covariates if clinically justified.

A standard margin of 0.80 to 1.25 for the ratio of geometric means for all PK parameters is suggested by regulatory guidelines[1,2] and accounts for an acceptable difference in systemic drug exposure between treatments of up to 20% (Figure 2).

For most products and indications no PD marker exists. In addition, when a PD marker does exist, the margin for the PD marker is highly dependent on the PD marker chosen and therefore needs to be defined for each compound individually and agreed with health authorities, following the same principles as for the efficacy margin in the confirmatory efficacy/safety study (see below). If a sensitive PD marker for the compound is available, efficacy can also be assessed in the PK/PD similarity study and may not have to be established in a confirmatory efficacy/safety study, which then would focus on safety and immunogenicity only. In any case, the EMA requires that at least 1 year of safety data be collected in the confirmatory efficacy/safety study.[3]

PK bridge and multiple comparisons

To demonstrate similar PK, three treatment comparisons are performed: biosimilar vs EU reference, biosimilar vs US reference, and the PK bridge of EU reference to US reference. The PK bridge, together with the analytical bridge (e.g., structural and functional data) comparing all three products (biosimilar, EU reference, and US reference), can then form the basis for justifying the relevance of data from in vivo non-clinical or clinical studies comparing the proposed biosimilar to a reference product authorised in a different region (for example using EU reference data for an FDA submission). This potentially reduces costs and development time by including only one reference product in animal studies or the confirmatory efficacy/safety study.[4,5]

Comparing all three products pairwise in the PK/PD similarity study leads to three treatment comparisons. In addition, multiple primary endpoints (AUC_{inf}, C_{max}, and AUC_{last}) may be assessed, leading to up to nine possible comparisons. As a 5% false positive rate (one-sided directional hypothesis) is inherent in all comparisons, counter-measures need to be taken to avoid an inflated rate of false positive conclusions. A number of methods are available for controlling the rate of false positive conclusions.[6] If multiple comparisons are made on multiple primary endpoints covering different aspects of the drug effect, all comparisons need to be successful for the study to be conclusive. As one option to control multiple comparisons, a hierarchical testing strategy can be applied, where all comparisons are first ranked in order and then each subsequent comparison is only tested if the previous higher-ranked comparison is successful. In this case, no adaption of the significance level for each individual comparison is required (Figure 3).

Another consideration for multiple comparisons is the impact on the power of the study. Studies are often powered at an overall level of 80%. Therefore, each individual comparison should be powered at a higher power (for example 96%) to ensure that the overall power of 80% is maintained.

Single active control confirmatory efficacy/safety study

The objective of the confirmatory efficacy/safety study is to demonstrate that no clinically meaningful differences exist between the proposed biosimilar and the reference product (active control) in terms of efficacy, safety, and immunogenicity. The objective of this study is not to demonstrate patient benefit per se, which has already been established for the reference medicinal product, and therefore no placebo arm is required. Instead, an indirect comparison to placebo should be made through estimation of the equivalence margin. Justification of the equivalence margin is based on the past performance of the reference product, often in the pivotal studies used for the reference product approvals. A systematic review is conducted to identify studies relevant to the comparison of the reference treatment versus placebo in the indication being considered. These studies can be used to estimate the effect size: difference between reference and placebo, with the corresponding CI. The planned confirmatory efficacy/safety study comparing the biosimilar product with the reference product will also provide an estimate of treatment effect with a CI. If these two CIs are combined, an indirect CI comparing the biosimilar test product and placebo can be obtained. Superiority of the biosimilar versus placebo is then demonstrated if the lower bound of the indirect CI is greater than zero (Figure 4).[7,8]

During biosimilar clinical development, similar efficacy and safety do not need to be demonstrated in every approved indication.

Most sensitive setting with regards to indication and primary endpoint

The objective of the confirmatory efficacy/safety study is to demonstrate that no clinically meaningful differences exist between the test and reference products in terms of efficacy and safety. Therefore, the comparison between the products needs to be performed using the most sensitive model (indication plus endpoint) and study conditions in a homogeneous patient population to detect any product-related differences, should they exist. The approved indication chosen is not necessarily the indication for which the product is most frequently used. The same most sensitive principle applies when selecting the primary endpoint for demonstrating similar efficacy. The chosen endpoint should be objective and exhibit a clear treatment effect. Often a continuous endpoint can be more sensitive to detect differences than a binary endpoint. For example, in an oncology setting overall survival and progression free survival are important clinical endpoints with which to establish patient benefit for a new anticancer drug. However, these endpoints may not be feasible or sensitive enough for assessing similarity of a proposed biosimilar and reference product since they may be influenced by various factors not attributable to differences between the biosimilar and the reference product, such as tumour burden, performance status, previous lines of treatment etc. Instead, overall response rate or percentage change in tumour mass from baseline may be used.[9]

During biosimilar clinical development, similar efficacy and safety do not need to be demonstrated in every approved indication. Instead, the confirmatory efficacy/safety study is conducted in the most sensitive indication only, and then biosimilarity is extrapolated to all other approved indications. A scientific argument including the mechanism of action of the product is provided to justify extrapolation to other indications and use of the full reference product label.

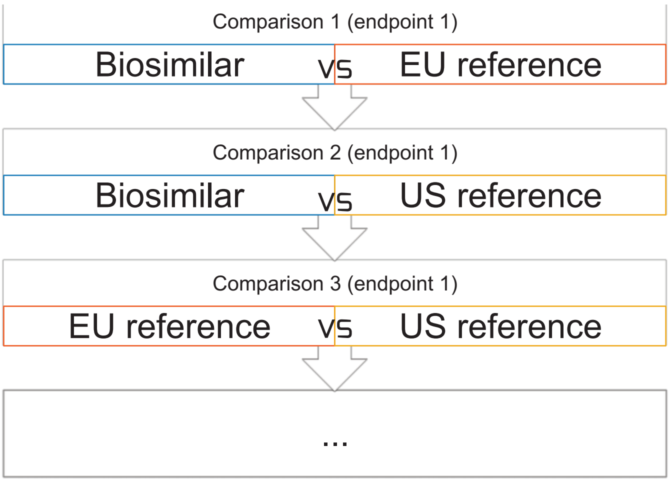


Figure 3. Hierarchical testing strategy. Three of the nine possible comparisons are shown.

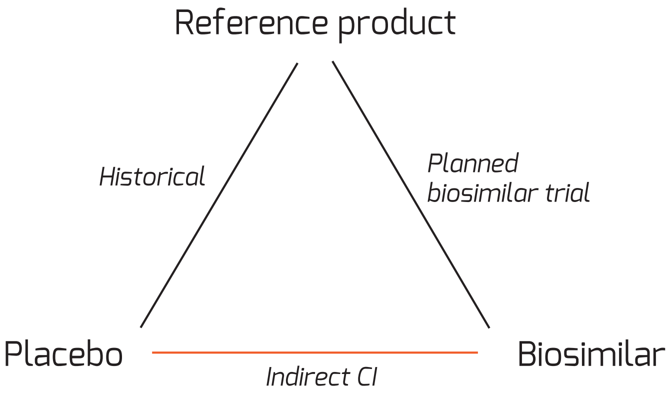


Figure 4. Indirect confidence interval (CI) showing superiority of biosimilar over placebo

Per-protocol analysis set as the most sensitive analysis set

Different to new study drug development where the objective is often to demonstrate superiority, for a biosimilar programme the objective is to demonstrate equivalence. Because of this, the primary statistical analysis of the primary endpoint should be performed on the per-protocol analysis set; the full analysis set is used for a secondary analysis. The per-protocol set is the cleanest analysis population to avoid biasing the comparison towards equivalence due to effect distortion by protocol deviations and imputation of missing data. Efficacy data for patients with major protocol deviations may not present an accurate picture of the product effect itself but are likely influenced by other factors. With such factors distorting the results for both the biosimilar and the reference drug it becomes increasingly difficult to detect any potential differences between actual product effects. This biases the comparison towards equivalence.

Analysis of the primary endpoint

Statistically, the comparison between the biosimilar and the reference product in terms of efficacy is performed by demonstrating that the 90% (for the FDA) or 95% (EMA) CI for the difference between the products for the primary endpoint falls entirely within a pre-defined margin. Figure 5 illustrates an example where PASI75 response rate (percentage of patients achieving a 75% reduction in Psoriasis Area and Severity Index) is the primary endpoint, the difference between test (biosimilar) and reference (EU reference) products is estimated as a risk difference, and the pre-defined margin is -18.0% to +18.0%. In this scenario, equivalent efficacy would be demonstrated by a risk difference of -2.5% with a 95% CI of -10.0% to 5.1% – or any other CI falling entirely inside the margin.

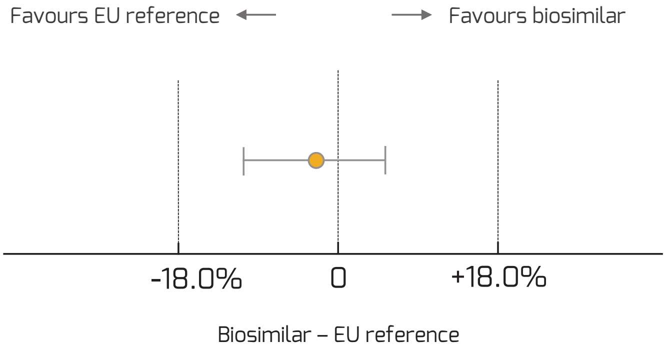


Figure 5. Plot for the risk difference between a biosimilar and an EU reference for PASI75 response rate. Equivalence met: 95% confidence interval for risk difference contained entirely within margin of -18.0% to +18.0%.

Importance of equivalence margin justification

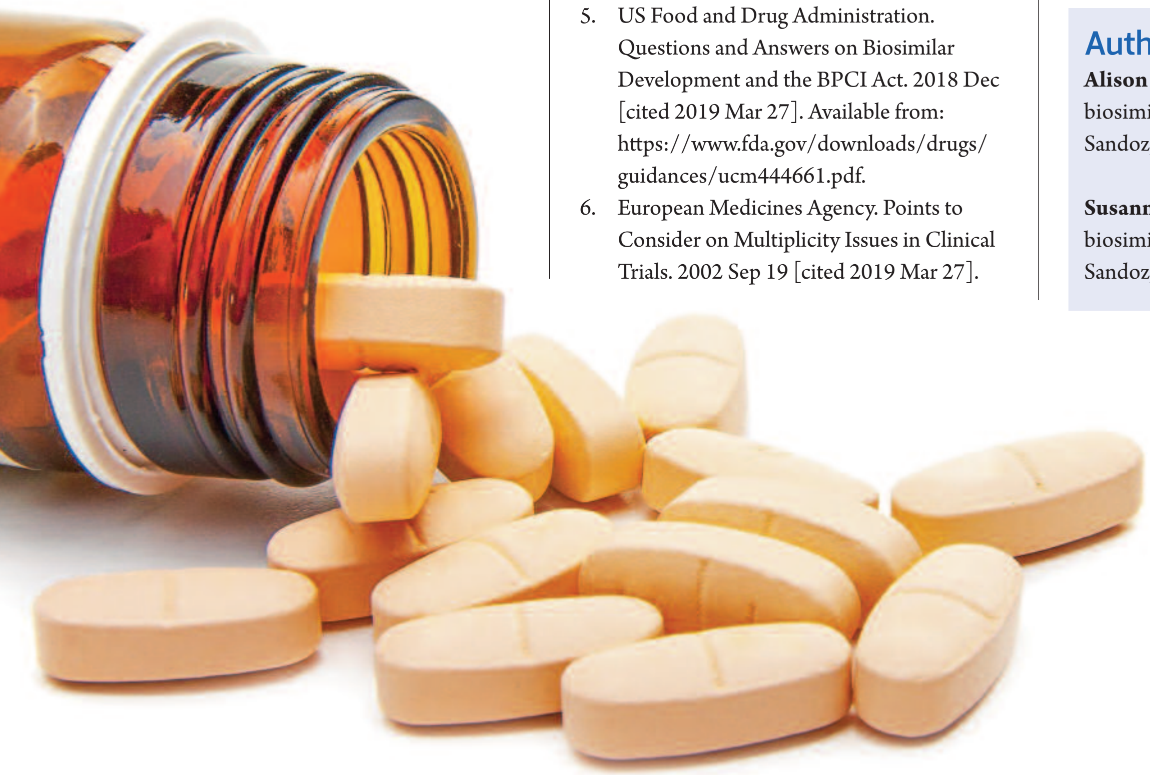
As the equivalence margin defines the equivalence criteria and also drives the study sample size, it needs to be selected carefully and agreed with health authorities. The margin is based on statistical as well as clinical considerations. Statistical significance pertains to whether or not the observed result could occur by chance alone, while clinical significance pertains to whether or not the observed result has "important" clinical, research, or public health relevance. The margin is derived based on past performance of the reference drug compared to placebo to ensure that the biosimilar drug maintains an agreed upon proportion (usually 50% or more) of the effect size of the reference drug. The effect size is estimated as the lower bound of the 95% CI for the difference between reference drug and placebo. A meta-analysis is performed where multiple data sources are available.[7, 8]

Potential bias towards equivalence when imputing missing data

When comparing the biosimilar and the reference drug, special considerations have to be given to the occurrence and imputation of missing data so as to not bias the results to equivalence. To counter this potential effect, the main analysis is usually based on the per-protocol set, thereby excluding patients with missing data. The robustness of the conclusion from the per-protocol set should be assessed through sensitivity analyses to account for different missing data scenarios. For imputation of missing data for both the biosimilar and the reference product using the same imputation rule, equivalence may be falsely concluded due to the imputation rather than a similar therapeutic effect.[10] For example, imputing all missing values as non-responders would reduce the treatment effect for both products and thereby reduce the treatment difference. One possibility is to impute the missing data for the reference product as responders and the missing data for the proposed biosimilar as non-responders and vice versa (extreme case scenarios). Alternatively, a tipping point analysis could be performed to understand the possible impact of missing data and which scenarios for the missing data would 'tip' the statistical analysis to no longer demonstrate equivalent efficacy.

Conclusion

With increased efforts to reduce health care costs, biosimilars have become more and more relevant. However, with biosimilars as a somewhat new concept in the world of medicinal product development, the regulatory environment and public understanding and acceptance are still evolving. As more guidelines on how to plan biosimilar trials become available, medical writers need to work closely with statisticians to determine which concepts from new drug development can be applied to biosimilar development and which aspects require different approaches. In addition, biosimilar-specific topics such as interchangeability [see Biosimilar development – an overview, p. 20] are still under discussion, making biosimilar development an interesting and highly relevant field to work in.



Conflicts of interest

The authors work for Sandoz, a Novartis division.

References

1. European Medicines Agency. Guideline on the Investigation of Bioequivalence. 2010 Jan 29 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/investigation-bioequivalence.
2. US Food and Drug Administration. Statistical Approaches to Establishing Bioequivalence. 2001 Jan [cited 2019 Mar 27]. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070244.pdf.
3. European Medicines Agency. Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Active Substance: Non-clinical and Clinical Issues. 2015 Jan 09 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/similar-biological-medicinal-products-containing-biotechnology-derived-proteins-active-substance-non.
4. European Medicines Agency. Guideline on similar biological medicinal products. 2014 Oct 23 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-similar-biological-medicinal-products-rev1_en.pdf.
5. US Food and Drug Administration. Questions and Answers on Biosimilar Development and the BPCI Act. 2018 Dec [cited 2019 Mar 27]. Available from: https://www.fda.gov/downloads/drugs/guidances/ucm444661.pdf.
6. European Medicines Agency. Points to Consider on Multiplicity Issues in Clinical Trials. 2002 Sep 19 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issues-clinical-trials_en.pdf.
7. European Medicines Agency. Guideline on the Choice of the Non-Inferiority Margin. 2005 Jul 27 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/choice-non-inferiority-margin.
8. US Food and Drug Administration. Non-Inferiority Clinical Trials to Establish Effectiveness. 2016 Nov [cited 2019 Mar 27]. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf.
9. European Medicines Agency. Guideline on Similar Biological Medicinal Products Containing Monoclonal Antibodies – Non-clinical and Clinical Issues. 2012 Jun 15 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/similar-biological-medicinal-products-containing-monoclonal-antibodies-non-clinical-clinical-issues.
10. European Medicines Agency. Guideline on Missing Data in Confirmatory Clinical Trials. 2010 Sep 20 [cited 2019 Mar 27]. Available from: https://www.ema.europa.eu/en/missing-data-confirmatory-clinical-trials.

Author information

Alison Balfour has been working on biosimilar trials as a Biostatistician within Sandoz, a Novartis division, since 2014.

Susanne Schmitt has been working on biosimilar trials as a Biostatistician within Sandoz, a Novartis division, since 2014.