Statistical Primer on Biosimilar Clinical Development

Leah Isakov, PhD, MBA,[1]* Bo Jin, PhD,[1] and Ira Allen Jacobs, MD, MBA, FACS[2]

A biosimilar is highly similar to a licensed biological product and has no clinically meaningful differences between the biological product and the reference (originator) product in terms of safety, purity, and potency and is approved under specific regulatory approval processes. Because both the originator and the potential biosimilar are large and structurally complex proteins, biosimilars are not generic equivalents of the originator. Thus, the regulatory approach for a small-molecule generic is not appropriate for a potential biosimilar. As a result, different study designs and statistical approaches are used in the assessment of a potential biosimilar. This review covers concepts and terminology used in statistical analyses in the clinical development of biosimilars so that clinicians can understand how similarity is evaluated. This should allow the clinician to understand the statistical considerations in biosimilar clinical trials and make informed prescribing decisions when an approved biosimilar is available.

Keywords: biosimilar, statistical analyses, equivalence

INTRODUCTION

A biosimilar is a biological therapy, which is highly similar to a licensed biological product (ie, originator biologic) and approved under a specific regulatory approval process, such as those enacted by the European Medicines Agency (EMA), US Food and Drug Administration (FDA), and the World Health Organization.[1–3] According to the definition in the FDA guidelines, a biosimilar has "no clinically meaningful differences between

the biological product and the reference (originator) product in terms of safety, purity, and potency."[2]

Biosimilars are not generic equivalents of the originator because both the originator and the potential biosimilar are large and structurally complex proteins, which can undergo posttranslational modifications that may affect safety and potency (even if the differences are minor).[2] Thus, although regulatory approval of small-molecule generic medicines can occur after demonstration of bioequivalence to the originator small-molecule drug, this approach is not sufficient for approval of potential biosimilars.[3] Instead, biosimilars are evaluated in a stepwise fashion that includes structural and functional studies, nonclinical assessments of pharmacokinetics (PK) and toxicity, and clinical evaluation of PK, efficacy, and safety (including immunogenicity) to demonstrate similarity of the potential biosimilar to the originator biologic.[1–4] Approval of a potential biosimilar is based on the totality of the evidence in demonstrating similarity of the biosimilar to the reference product.[1–4] Totality of the evidence refers to the extensive comparative structural, functional, nonclinical, and clinical data required to establish biosimilarity. Totality of the evidence is an important concept for biosimilars because of the different paradigm for developing and regulating biosimilar products. Data from all stages of development

www.americantherapeutics.com

(not just clinical) are important for regulatory approval and to gain medical acceptance.

The number of biosimilars approved by regulatory agencies is expected to increase as patents on licensed originator biologics expire. Thus, it is important for prescribers to understand statistical analyses used in biosimilar clinical trials so that they can make informed decisions about appropriate ways to incorporate biosimilars into treatment regimens. The purpose of this review is to provide a high-level education to clinicians on concepts and terminology used in statistical analyses in the clinical development of biosimilars to understand how similarity is evaluated. Two case studies are also presented to illustrate margin derivation and sample size calculation for efficacy equivalence trials for potential biosimilars.

TERMINOLOGY AND DEFINITIONS

Clinicians should become familiar with the traditional statistical approaches used in approval process for evaluation of a novel therapeutic (eg, the New Drug Application or the Biologics License Application processes), which generally use *superiority trial* study designs. A superiority trial has a primary objective of showing response to the investigational product is superior to a comparative agent (active or placebo control).[5] In many cases, the comparative agent is the current standard of care, especially if treatment with a placebo would be considered unethical.

In contrast, biosimilar products must demonstrate similar efficacy and safety to the originator. Typically, studies of a potential biosimilar usually use an *equivalence trial* design.[2,3] An equivalence trial has a primary objective of showing that differences in response to 2 or more treatments are clinically unimportant.[5] In other words, clinical responses are close enough so neither the biosimilar nor the comparator (originator) is superior or inferior to the other.[6] This is usually shown via demonstration that the true treatment difference is likely to lie within a specific range of clinically acceptable differences.[5] Equivalence is demonstrated when the entire confidence interval (CI), or a given parameter, falls within the lower and upper equivalence margins set before the experiment or study (Figure 1).

Although the equivalence trial is commonly used, some assessments of a biosimilar may use a *noninferiority trial* design if adequate scientific justification is provided (Figure 2).[2,3] In a noninferiority trial, the primary objective is to show that response to the investigational product is not clinically inferior to a comparative agent (active or placebo control).[5] A noninferiority design uses only one margin (the lower or upper limit,
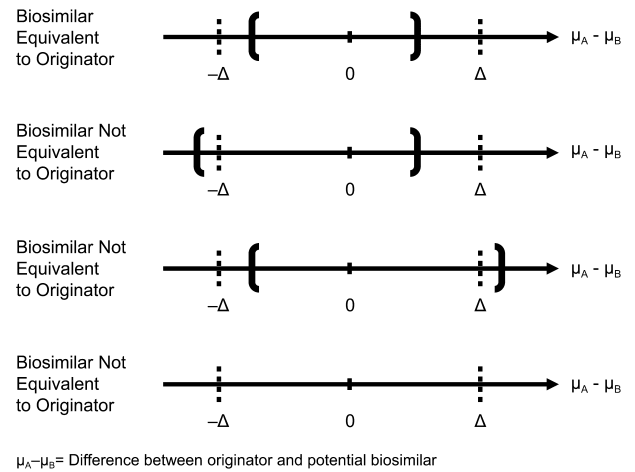

FIGURE 1. Testing for equivalence of the potential biosimilar to the originator biologic.

depending on what is appropriate for the specific study or endpoint) and tends to require a smaller sample size than an equivalence trial design.[3]

The CI expresses the degree of uncertainty associated with a statistical parameter, such as the difference between 2 treatment effects (such as risk difference) or the ratio. For example, if overall response rate (ORR) is used as the primary endpoint, risk ratio could serve as a primary efficacy parameter. The CI is different from the *point estimate* for a population parameter. For example, sample mean, ORR, and median survival time are examples of point estimates of unknown population parameters.

The *margins* for demonstration of equivalence (and noninferiority) are generally based on the effect size, should be justified on both clinical and statistical grounds, and use robust statistical rationale and clinical criteria to determine a margin value.[7] In terms of demonstrating biosimilarity, the margin is the largest difference between the potential biosimilar and originator that is judged as clinically acceptable (Table 1) but should be smaller than the minimum difference reported between the originator and a placebo for clinically relevant results.[7,8] When available, effect size, magnitude, and
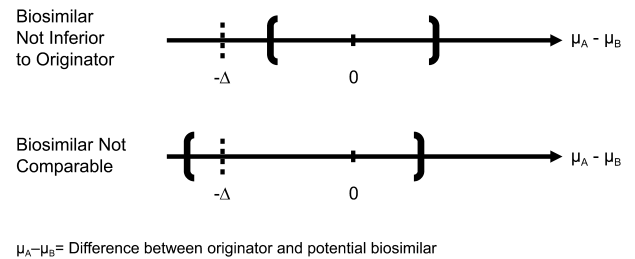

FIGURE 2. Testing for noninferiority of the potential biosimilar to the originator biologic.

Table 1. Sample size calculations under different scenarios.

Power	ORR	Sample size	Sample size after accounting for 10% dropout
Equivalence margin (0.73, 1.38)			
0.80	0.38	574	638
0.85	0.38	636	707
0.90	0.38	724	804
0.80	0.40	528	587
0.85	0.40	586	651
0.90	0.40	666	740
Equivalence margin (0.75, 1.34)			
0.80	0.38	688	764
0.85	0.38	764	849
0.90	0.38	870	967
0.80	0.40	634	704
0.85	0.40	704	782
0.90	0.40	800	889

variability of effect size of the originator in placebo-controlled trials may be used to calculate margins.[7,8] In some cases, such as in clinical PK studies, which lack established acceptance criteria for biologicals, regulators suggest a traditional equivalence range of 80%–125%.[3]

The *two one-sided test* (TOST) is the simplest and most widely used approach to test equivalence.[4,6] This is done using 2 simultaneous 1-sided tests to analyze the composite null and alternative hypotheses (H_0 and H_1, respectively), which are: H_0: $\mu \leq \theta_L$ or $\mu \geq \theta_U$ (the treatment difference is outside the equivalence margin) and H_1: $\theta_L < \mu < \theta_U$ (the treatment difference is within the margins), where θ_L and θ_U are the lower and upper margins, respectively, and μ is the analysis criterion (eg, mean ratio or mean difference).[4,9,10] This is done with 2 sets of hypotheses (H_{a0}: $\mu \leq \theta_L$ with H_{a1}: $\mu > \theta_L$ and H_{b0}: $\mu \geq \theta_U$ with H_{b1}: $\mu < \theta_U$).[4,9] The H_0 is rejected in favor of H_1 if the CI for μ is contained completely within the range (θ_L–θ_U).

In any statistical test, 2 types of errors can result: *type I error* or *type II error*. A type I error occurs when the null hypothesis is rejected when it is true, whereas a type II error occurs when the null hypothesis is accepted when it is false.[11] The probability of committing a type I error (denoted by α) is the significance level for the test (eg, α = 0.05), whereas the probability of committing a type II error is denoted by β.[11] The probability of not

committing a type II error is the *power* of the test, which is equal to 1 minus β.[11] In the context of the TOST procedure, the type I error is the larger of the 2 sets of hypothesis tests and the power for the TOST procedure is the probability of not committing the type II error in either of the 2 sets of hypotheses tested.

STATISTICAL CONCEPTS FOR CLINICAL STUDIES OF BIOSIMILARS

Biosimilars are evaluated in a stepwise fashion requiring demonstration of similar structural, functional, and nonclinical PK and nonclinical toxicity to the originator before clinical trials begin.[1–4] The specific nature of comparative clinical trials depends on the nature and extent of residual uncertainty about biosimilarity.[2] The EMA, FDA, and World Health Organization guidelines for biosimilar development note that clinical PK is a critical basic characteristic of any medicinal product and should be evaluated as the first step of any clinical program.[1–3] These studies should be comparative in nature and designed to enable detection of potential differences between the potential biosimilar and the originator.[3] If feasible or if relevant pharmacodynamic measures are available, initial clinical studies may include pharmacodynamics in addition to PK.[1,2]

Although the exact nature of statistical analyses depends on trial design, there are some principles that generally apply in most clinical studies evaluating potential biosimilars. A single-dose crossover study is not appropriate for most biosimilar clinical PK studies because of the long half-life and the potential for formation of anti-drug antibodies. Instead, parallel treatments are necessary.[3] This means that to show bioequivalence, sample size needs to be increased relative to crossover studies.[3]

Regulatory agencies generally expect that equivalence trials will be used unless there is sufficient scientific justification for using noninferiority designs.[1] For equivalence trials, statistical analysis is generally based on the use of 2-sided CIs (typically at the 90% level) for the difference between treatments.[3] In many situations, no established acceptance criteria exist for standard clinical PK comparability (bioequivalence) studies for biological drugs because the criteria were developed for chemically derived orally administered products and, thus, may not be applicable for biologics.[3] In this case, regulatory agencies often recommend use of the traditional 80%–125% equivalence range with 90% CIs of the ratio of the population geometric means (test to reference) for studies evaluating clinical PK.[3,12] However, this is not the default range and proposed limits need to be justified.[13]

Regardless, similarity limits are determined before initiation of studies.

Outcomes can be measured as the treatment effect, which is either an absolute measure or a relative measure.[6,8] For example, the *risk difference* is an absolute measure of effect and is calculated by subtracting the risk of the outcome for 2 groups of individuals (eg, those treated with the potential biosimilar and those treated with the originator). The *relative risk* (RR) (also called *risk ratio*) is a relative measure of effect calculated as the ratio between 2 incidence proportions and is generally presented with the CI as a measure of precision of the estimate.[14] According to the FDA Statistical Approaches to Establishing Bioequivalence Guidance for Industry, the primary comparison in bioequivalence studies is expected to be the ratio between the average data for each parameter (eg, area under the curve) between the test (biosimilar) and the reference (originator), which can be conducted as a general comparison via logarithmic transformation, rather than the difference between the averages.[15]

The patient population evaluated in clinical trials is also important. The intent-to-treat population includes every subject who is randomized and does not remove patients with noncompliance, protocol deviations, withdrawal, or anything that happens after randomization from the analysis.[16] Although considered a conservative analysis in the superiority trial setting and therefore appropriate for the primary analysis, the intent-to-treat population is not conservative in equivalence and noninferiority trials because of a tendency for biasing the results toward equivalence (or noninferiority).[8] However, the per-protocol population, which excludes data from patients with major protocol violations, can also bias results.[8,16] Therefore, both populations should be evaluated and reported in equivalence (and noninferiority) trials.[8] It is expected that at the end of the trial analysis, both intent-to-treat and per-protocol populations lead to the same conclusion for potential biosimilars.

STATISTICS IN COMPARATIVE EFFICACY AND SAFETY STUDIES

After demonstration of comparable clinical PK, the next step in the development program is usually to evaluate the clinical efficacy and safety in adequately powered and randomized controlled clinical trial(s).[1,3] The goal of these studies is to demonstrate that there are no clinically meaningful differences in efficacy, safety, and immunogenicity of the potential biosimilar compared with the originator.[1–3] These studies should include a sensitive patient population and endpoints with appropriate scientific justification.[1–3]

For comparisons of efficacy, selection of an appropriate primary endpoint depends on several factors, including the originator, intended use of the potential biosimilar, disease prevalence, and target population.[3] The primary endpoint is expected to be relevant and sensitive to detect potential differences between the potential biosimilar and originator.[1] This means that choice of clinical endpoints and/or time points for analyses may differ for those used for regulatory approval of the originator or for those traditionally applied for a novel agent in the same therapeutic area.[1,2] When possible, some secondary endpoints used for approval of the originator should also be included to facilitate comparisons between drugs.[1,2] Data collection for secondary endpoints may require an additional follow-up period. Although some regulatory agencies have provided some class-specific guidelines defining appropriate endpoints for biosimilars, not all areas are covered by these guidelines.[1] One approach to identifying sensitive populations, margins, and appropriate endpoints in clinical efficacy is to perform a meta-analysis of clinical data for the originator.[17] Regardless of the endpoints selected, equivalence designs are generally preferred for comparisons and margins must be prespecified and include the largest differences that would not be clinically relevant.[2,3] Generally, the prespecified margins are symmetric, although an asymmetric margin can be used if dose-related toxicities occur or if the dose used is near the plateau of the dose–response curve and there is little likelihood of dose-related effects.[2]

Although investigation of safety and tolerability is multidimensional because of the wide range of possible adverse effects and that occurrence of new and unforeseeable effects is always possible, safety analyses for potential biosimilars are expected to compare severity and frequency of adverse events after administration of the potential biosimilar versus the originator, with specific interest in adverse events listed in the originator product label.[1–3,5] In addition, a key element of biosimilar safety is assessment of potential differences in incidence and severity of human immune responses because these may affect safety and effectiveness.[1–3] Recommendations for statistical analyses of safety and tolerability in clinical trials include descriptive statistics of the data with provision of CIs and/or *P* values when useful for interpretation.[5]

CASE STUDIES: ILLUSTRATING MARGIN DERIVATION AND SAMPLE SIZE CALCULATION FOR EFFICACY EQUIVALENCE TRIALS

As described earlier, biosimilar products must demonstrate comparable efficacy and safety to the originator

rather than superiority. This is usually done using an equivalence trial design with appropriate margins and sample sizes to show that the treatment difference lies within the calculated range of clinically acceptable differences. The calculation of equivalence margins and sample size depends on the selection of appropriate and sensitive clinical endpoints, and this, in turn, depends on the therapeutic indication of interest. Two case studies illustrate the approaches used when designing clinical trials to evaluate a potential biosimilar with indications for cancer or inflammatory diseases.

Case study 1: oncology

In general, the amount of clinical trial data supporting registration of monoclonal antibodies used in the treatment of cancer is limited. Instead of being developed as a novel monotherapy for the treatment of specific types of cancer, in general, monoclonal antibodies are commonly evaluated as "add-on" therapy to a current, cytotoxic chemotherapy regimen used as the standard of care. It is rare to find large randomized controlled clinical trials that all use the same patient population and cytotoxic chemotherapy. Although randomized trials may have been conducted, they are often smaller studies, in different patient populations, and/or with different chemotherapy regimens. This leads to considerable variability in reported outcomes and inconsistent treatment effects. These differences pose a challenge for the identification of a robust and consistent treatment effect margin that could be used in the design of a biosimilar equivalence study.

According to the FDA and EMA biosimilar guidelines, selection of the most sensitive population is critical step in the design of efficacy trial.[1,2] Therefore, the selection of an appropriate study population usually occurs after discussions with regulators and is based on historical studies, practical considerations, and input from clinicians to address changes in clinical practice and current knowledge of the disease.[1,2] Whenever possible, the potential for confounding

because of co-medication or previous treatment must be avoided. Thus, the selection of the sensitive patient population may be based on an indication in which the originator is approved for first-line use (with or without concomitant chemotherapy). For example, patients with previously untreated, advanced non–small-cell lung cancer (NSCLC) are well characterized for safety and efficacy and therefore are considered a sensitive population for the evaluation of potential biosimilars to bevacizumab (in combination with paclitaxel and carboplatin) per the bevacizumab-approved indications in the United States and European Union.[17–19]

Identifying appropriate endpoints for comparisons between a potential biosimilar and the originator must take into consideration how to identify any potential product-related differences in efficacy and safety while removing potential confounding factors.[20] Thus, although survival endpoints are preferred in clinical trials evaluating novel biological therapies for oncology indications, these endpoints may not be appropriate in clinical trials evaluating biosimilars. Instead, a measure of response (such as ORR) or activity may be more sensitive for clinical comparisons of a potential biosimilar and therefore a better choice for the primary endpoint.[1,20,21]

When an appropriate study population and primary endpoint are identified, the next step is to derive the equivalence margin to demonstrate biosimilarity between the potential biosimilar and the originator. This case study is a hypothetical phase 3 comparative clinical trial of a potential biosimilar to bevacizumab being evaluated for safety and efficacy in a study population of patients with NSCLC and a primary endpoint of ORR. The equivalence margin can be calculated based on the meta-analysis described earlier that evaluated the addition of bevacizumab to chemotherapy in patients with NSCLC (Figure 3).[17] Margin construction is commonly designed to preserve at least 50% of the treatment effect.[10] In this example, using the 4 clinical trials identified as appropriate for this patient population and primary endpoint,[22–25] we

Study	Events	RR Total	Events	Control Total	Risk Ratio	RR	[70% CI]	W(fixed)	W(random)
Johnson (2004)	11	34	6	32		1.73	[0.72, 4.12]	1.8%	2.2%
Niho (2012)	68	121	20	59		1.66	[1.12, 2.45]	8.9%	10.4%
Reck (2010)	114	329	71	327		1.60	[1.24, 2.06]	21.1%	22.6%
Sandler (2006)	133	381	59	392		2.32	[1.77, 3.05]	18.3%	20.0%
	326	865	156	810		1.96	[1.66, 2.31]	49.9%	44.8%
Fixed effect model	1730		1620			1.90	[1.79, 2.02]	100%	--
Random effects model						1.89	[1.77, 2.03]	--	100%

Heterogeneity: I-squared=11.5%
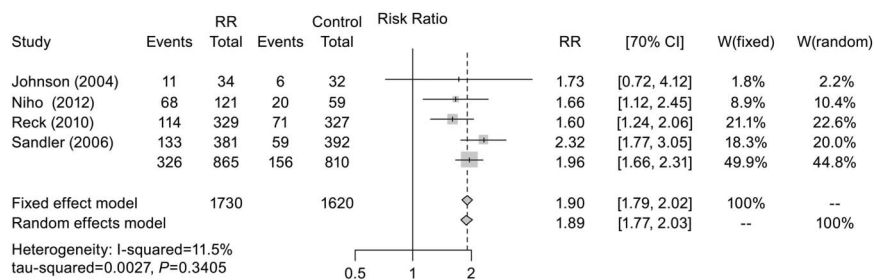tau-squared=0.0027, P=0.3405

0.5   1   2

FIGURE 3. Risk ratio in ORR for calculating the equivalence margin needs to compare a potential biosimilar for bevacizumab to the originator.

applied log scale and our margins are symmetrical around 1. Log scale transformation is commonly used when a ratio (eg, RR or hazard ratio) serves as a primary endpoint. Therefore, computing the margins for a risk ratio is a 2-step procedure. First, the upper margin is generated for Ln(RR) and then exponent of the upper margin is computed to give the upper limit margin for the RR. Then to construct a lower margin that is symmetrical, the reciprocal of the upper margin is taken. Based on a log scale, computing margins for Ln(RR): 1/2 of Ln(1.79) = 0.291, the upper margin is exp(0.291) = 1.34 and lower margin is 1/1.34 = 0.75. Thus, an equivalence margin of 0.75, 1.34, will demonstrate clinical similarity for the intended indication while maintaining 50% of the 2-sided 70% CI lower bound of the effect size based on the historical ORR data.

Treatment effect is a critical component for sample size calculation. In this example, the fixed-effect model estimates that the ORR (95% CI) for chemotherapy plus bevacizumab is 38% (35%, 40%; Figure 4). This relatively conservative assumption of an ORR of 38% is used to calculate the total sample size. If the type I error rate is 5% and power is 85%, the total sample size would be 764 (without accounting for dropout; Table 1). Under an assumption of 10% drop-out rate, the target sample size would be 849 patients. Although other assumptions about power will change the sample size (eg, using power of 80% results in a decreased total sample size of 688 patients and increasing power to 90% leads to an increase in the required sample size to 870 patients), the estimate of the treatment effect has a greater impact on sample size. For example, if the assumed ORR is 40% (which is potentially justifiable based on the fact that this value is the upper limit of the 95% CI for the fixed-effect model), the sample size will be 633 patients for power of 80%, 704 patients at power of 85%, and 800 patients if the power used is 90% without accounting for dropout. Using a 45% preservation of treatment effect (instead of the standard preservation of 50% of treatment effect) will lead to wider margins and, as a result, to smaller sample

size under the same set of assumptions. It is important to remember that other approaches for treatment effects and sample size calculations could apply if clinical interpretations suggest that it is more appropriate to use nonsymmetrical margins, linear scale, or noninferiority margins.

Case study 2: rheumatoid arthritis

Clinical efficacy of a therapeutic agent used in the treatment of inflammatory diseases, such as rheumatoid arthritis, can be evaluated through a number of measures, including composite measures of disease activity. In this case study, a hypothetical study of a potential rituximab biosimilar (with methotrexate) being evaluated for safety and efficacy in a study population of patients with rheumatoid arthritis, the primary endpoint selected is the disease activity in 28 joints score (DAS28). DAS28 is a validated, continuous composite measure of disease activity included as an endpoint in clinical trials in rheumatoid arthritis, including those evaluating rituximab plus methotrexate.[26–32] DAS28 is considered a sensitive continuous endpoint to detect potential differences in treatment effect in a potential biosimilar.[27,33]

Using a meta-analysis approach similar to one used by Volkmann et al to assess clinical efficacy, an equivalence margin for DAS28 in terms of change from baseline at week 24 can be derived (see Volkmann et al, Figure 3).[32] The FDA guidance for noninferiority trials suggests that a margin for noninferiority trial preserves at least 50% efficacy from historical data.[10] In this case, the noninferiority margin would be 50% of the upper 95% CI, (−0.70) for the treatment effect difference (vs. control) from historical data (ie, −0.35) and the symmetric margin for an equivalence trial is then −0.35, 0.35.

Using this margin of −0.35, 0.35, and assuming that the true difference is zero, a sample size of 356 patients will provide 90% power to establish equivalence with type I error of 0.05 in terms of standardized DAS28 change from baseline at week 24. If a 20% attrition rate
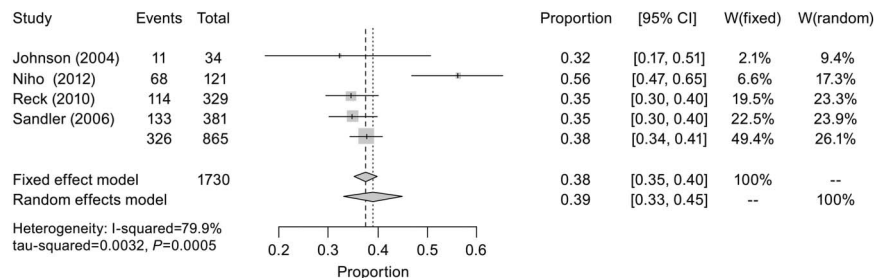

Study	Events	Total		Proportion	[95% CI]	W(fixed)	W(random)
Johnson (2004)	11	34		0.32	[0.17, 0.51]	2.1%	9.4%
Niho (2012)	68	121		0.56	[0.47, 0.65]	6.6%	17.3%
Reck (2010)	114	329		0.35	[0.30, 0.40]	19.5%	23.3%
Sandler (2006)	133	381		0.35	[0.30, 0.40]	22.5%	23.9%
	326	865		0.38	[0.34, 0.41]	49.4%	26.1%
Fixed effect model		1730		0.38	[0.35, 0.40]	100%	--
Random effects model				0.39	[0.33, 0.45]	--	100%

Heterogeneity: I-squared=79.9%
tau-squared=0.0032, P=0.0005

Proportion: 0.2 0.3 0.4 0.5 0.6

FIGURE 4. Treatment effect calculation for ORR (with 95% CI) for bevacizumab plus chemotherapy using historic data.

by week 24 is used, the trial will need to enroll a total sample size of 445.

CONCLUSIONS

Statistical analyses used for clinical assessments of potential biosimilars are different than those used for regulatory approval of the originator or those applied for approval of a small-molecule generic drug. Understanding the different analyses, comparisons, endpoints, populations used, and calculations of margins and sample size in these studies should allow the clinician to make informed decisions when prescribing biological agents when a biosimilar is approved by relevant regulatory authorities.

REFERENCES

1. European Medicines Agency. *Guideline on Similar Biological Medicinal Products*. Available at: http://www.ema. europa.eu/docs/en_GB/document_library/Scientific_ guideline/2014/10/WC500176768.pdf. Accessed June 5, 2015.
2. U.S. Food and Drug Administration. *Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. Draft Guidelines*. U.S. Department of Health and Human Services Food and Drug Administration; 2015. Available at: http://www.fda.gov/downloads/Drugs/ GuidanceComplianceRegulatoryInformation/Guidances/ UCM291128.pdf. Accessed April 7, 2015.
3. World Health Organization. *Guidelines on Evaluation of Similar Biotherapeutic Products (SBPs)*. Available at: http:// www.who.int/biologicals/areas/biological_therapeutics/ BIOTHERAPEUTICS_FOR_WEB_22APRIL2010.pdf. Accessed April 7, 2015.
4. Tsong Y, Dong X, Shen M. Development of statistical methods for analytical similarity assessment. *J Biopharm Stat.* 2015 Sep 15. [Epub ahead of print].
5. U.S. Food and Drug Administration. *Guidance for Industry E9 Statistical Principles for Clinical Trials*. U.S. Department of Health and Human Services Food and Drug Administration; 1998. Available at: http://www.fda.gov/downloads/ drugs/guidancecomplianceregulatoryinformation/ guidances/ucm073137.pdf. Accessed July 1, 2015.
6. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. *J Gen Intern Med.* 2011;26:192–196.
7. Dasgupta A, Lawson KA, Wilson JP. Evaluating equivalence and noninferiority trials. *Am J Health Syst Pharm.* 2010;67:1337–1343.
8. Njue C. Statistical considerations for confirmatory clinical trials for similar biotherapeutic products. *Biologicals.* 2011;39:266–269.
9. SAS Institute Inc. *SAS/STAT 9.2 User's Guide. TOST Equivalence Test*. 2nd ed. Available at: http://support. sas.com/documentation/cdl/en/statug/63033/HTML/ default/viewer.htm#statug_ttest_a0000000128.htm. Accessed July 2, 2015.
10. United States Food and Drug Administration. *Guidance for Industry Non-Inferiority Clinical Trials. Draft Guidance*. Available at: http://www.fda.gov/downloads/ Drugs/…/Guidances/UCM202140.pdf. Accessed September 24, 2015.
11. SAS Institute Inc. *SAS/STAT 9.2 User's Guide [computer Program]. Version 2*. SAS Institute Inc., Cary, NC; 2010.
12. Chow SC, Liu JP. Statistical assessment of biosimilar products. *J Biopharm Stat.* 2010;20:10–30.
13. U.S. Food and Drug Administration. *Guidance for Industry Clinical Pharmacology Data to Support a Demonstration of Biosimilarity to a Reference Product. Draft Guidance*. U.S. Department of Health and Human Services Food and Drug Administration; 2014. Available at: http://www. fda.gov/downloads/drugs/guidancecomplianceregula toryinformation/guidances/ucm397017.pdf. Accessed July 1, 2015.
14. Tripepi G, Jager KJ, Dekker FW, et al. Measures of effect: relative risks, odds ratios, risk difference, and "number needed to treat". *Kidney Int.* 2007;72:789–791.
15. U.S. Food and Drug Administration. *Guidance for Industry Statistical Approaches to Establishing Bioequivalence*. U. S. Department of Health and Human Services Food and Drug Administration; 2001. Available at: http://www. fda.gov/downloads/drugs/guidancecomplianceregula toryinformation/guidances/ucm070244.pdf. Accessed July 1, 2015.
16. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res.* 2011;2:109–112.
17. Botrel TE, Clark O, Clark L, et al. Efficacy of bevacizumab (Bev) plus chemotherapy (CT) compared to CT alone in previously untreated locally advanced or metastatic non-small cell lung cancer (NSCLC): systematic review and meta-analysis. *Lung Cancer.* 2011;74:89–97.
18. Genentech Inc. *Avastin (Bevacizumab) Prescribing Information*. South San Francisco, CA. Available at: http:// www.gene.com/download/pdf/avastin_prescribing. pdf. Accessed September 4, 2015.
19. Roche. *Summary of Product Characteristics. Avastin (Bevacizumab)*. Available at: http://www.ema.europa.eu/docs/en_ GB/document_library/EPAR_-_Product_Information/ human/000582/WC500029271.pdf. Accessed September 4, 2015.
20. Socinski MA, Curigliano G, Jacobs I, et al. Clinical considerations for the development of biosimilars in oncology. *MAbs.* 2015;7:286–293.
21. European Medicines Agency. *Guideline on Immunogenicity Assessment of Biotechnology-Derived Therapeutic Proteins*. Available at: http://www.ema.europa.eu/docs/ en_GB/document_library/Scientific_guideline/2009/ 09/WC500003946.pdf. Accessed July 15, 2015.
22. Johnson DH, Fehrenbacher L, Novotny WF, et al. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. *J Clin Oncol.* 2004;22:2184–2191.

23. Niho S, Kunitoh H, Nokihara H, et al. Randomized phase II study of first-line carboplatin-paclitaxel with or without bevacizumab in Japanese patients with advanced non-squamous non-small-cell lung cancer. *Lung Cancer.* 2012;76:362–367.

24. Reck M, von Pawel J, Zatloukal P, et al. Overall survival with cisplatin-gemcitabine and bevacizumab or placebo as first-line therapy for nonsquamous non-small-cell lung cancer: results from a randomised phase III trial (AVAiL). *Ann Oncol.* 2010;21:1804–1809.

25. Sandler A, Gray R, Perry MC, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med.* 2006;355:2542–2550.

26. Cohen SB, Emery P, Greenwald MW, et al. Rituximab for rheumatoid arthritis refractory to anti-tumor necrosis factor therapy: results of a multicenter, randomized, double-blind, placebo-controlled, phase III trial evaluating primary efficacy and safety at twenty-four weeks. *Arthritis Rheum.* 2006;54:2793–2806.

27. Dougados M, Schmidely N, Le Bars M, et al. Evaluation of different methods used to assess disease activity in rheumatoid arthritis: analyses of abatacept clinical trial data. *Ann Rheum Dis.* 2009;68:484–489.

28. Edwards JC, Szczepanski L, Szechinski J, et al. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. *N Engl J Med.* 2004;350:2572–2581.

29. Emery P, Deodhar A, Rigby WF, et al. Efficacy and safety of different doses and retreatment of rituximab: a randomised, placebo-controlled trial in patients who are biological naive with active rheumatoid arthritis and an inadequate response to methotrexate (Study Evaluating Rituximab's Efficacy in MTX iNadequate rEsponders (SERENE)). *Ann Rheum Dis.* 2010;69: 1629–1635.

30. Emery P, Fleischmann R, Filipowicz-Sosnowska A, et al. The efficacy and safety of rituximab in patients with active rheumatoid arthritis despite methotrexate treatment: results of a phase IIB randomized, double-blind, placebo-controlled, dose-ranging trial. *Arthritis Rheum.* 2006;54:1390–1400.

31. Fransen J, van Riel PL. The disease activity score and the EULAR response criteria. *Clin Exp Rheumatol.* 2005;23: S93–S99.

32. Volkmann E, Agrawal H, Maranian P, et al. Rituximab for rheumatoid arthritis: a meta-analysis and systematic review. *Clin Med Insights.* 2010;2:749–760.

33. Gaujoux-Viala C, Mouterde G, Baillet A, et al. Evaluating disease activity in rheumatoid arthritis: which composite index is best? A systematic literature analysis of studies comparing the psychometric properties of the DAS, DAS28, SDAI and CDAI. *Joint Bone Spine.* 2012; 79:149–155.