Therapeutic Innovation
& Regulatory Science
2018, Vol. 52(3) 300-305
© The Author(s) 2017
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2168479017729189
tirs.sagepub.com

Sample Size for Biosimilar Trials: In Defense of Synthesis

Timothy Clark, BSc DipStat, PhD[1], Sook Jung Jo, BSc, PhD[2], and Alan Phillips, BSc, PhD[2]

Abstract

Biosimilars are biological products similar to, but not the same as, the innovator products. Both the European Medicines Agency and the Food and Drug Administration have released detailed guidance on the development of biosimilars. This guidance requires the pivotal phase 3 clinical study to have an equivalence design, which means that the study objective is to demonstrate that one treatment is neither "worse than" nor "better than" the other by some "clinically unimportant" amount. The most critical and controversial step in designing such a study is the choice of equivalence margin, as this determines the conclusion of the study. In this paper, we outline the methodology for determining an equivalence margin and, through case studies on biosimilar trastuzumab (HERCEPTIN) and biosimilar bevacizumab (AVASTIN), explain the challenges of applying this in practice and why the synthesis method should be given greater consideration by regulatory authorities and biosimilar developers.

Keywords

biosimilar, equivalence margin, sample size, reference product

A biosimilar program usually consists of a phase 1 study to demonstrate pharmacokinetic similarity (area under the curve [AUC] and maximum observed concentration [Cmax]) and a phase 3 study to demonstrate comparable efficacy, safety, and immunogenicity. For the latter, the European Medicines Agency (EMA) and the Food and Drug Administration (FDA) both stipulate that an equivalence design is required.[1,2]

In an equivalence design, the goal is to show that two treatments are therapeutically equivalent, that is, that the difference between the treatments lies within a predefined equivalence margin.[3] Two treatments would be considered equivalent if the upper and the lower bounds of the confidence interval of the observed difference does not cross the prespecified margin.[4] Normally 95% confidence limits are used, although in bioequivalence trials it is the norm to use 90% limits.[4]

Two of the most commonly used methods to selecting the margin are the fixed margin (or 95-95) approach and the synthesis method (also known as the putative placebo method).[4-7] Both methods are outlined in the draft FDA guideline on non-inferiority, which was published in March 2010.[7]

Finally, it should be mentioned for completeness that Bayesian approaches can yield reduced sample sizes, but these will not be specifically addressed in this paper.[8]

Fixed Margin

The hypotheses for the fixed margin approach can be expressed as follows:

$$H_0\colon \mu_T - \mu_C \le -\Delta \text{ or } \mu_T - \mu_C \ge +\Delta \text{ versus } H_1\colon -\Delta < \mu_T - \mu_C < +\Delta,$$

where $-\Delta$ (≤ 0) and $+\Delta$ (≥ 0) are the prespecified fixed margins.[9]

Equivalence is concluded when each one-sided null hypothesis is rejected or, equivalently, if the $1 - 2\alpha$ (eg, 90%) two-sided confidence interval of the treatment difference, $\mu_T - \mu_C$, lies entirely inside $[\pm\Delta]$.

According to the FDA guidance, the first step in the fixed-margin method is to define the largest acceptable margin (M1) and the clinical margin (M2), which are key parameters for the sample size calculation. M1 is the effect of the active control, which is an assumed value based on the analysis of the effect of the active control seen in past controlled studies. M2 reflects the clinical judgment about how much of M1 should be preserved by ruling out a loss of M2.

[1] Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Faculty of Medicine, Ludwig-Maximilians University, Munich, Germany
[2] ICON Clinical Research, Seoul, Republic of Korea

Submitted 23-Mar-2017; accepted 7-Aug-2017

Corresponding Author:
Timothy Clark, BSc, DipStat, PhD, Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Faculty of Medicine, Ludwig-Maximilians University, Munich, Germany.
Email: Tim.Clark@lmu.de

The determination of M2 is based on clinical judgment and is usually calculated by taking a percentage or fraction of M1. The clinical judgment in determining M2 may take into account the actual disease incidence or prevalence and its impact on the practicality of sample sizes that would have to be accrued for a study. There can be flexibility in the M2 margin, for example, when the difference between the active comparator response rate and the spontaneous response rate is large or the primary endpoint does not involve an irreversible outcome such as death (in general, the M2 margin will be more stringent when treatment failure results in an irreversible outcome).

Synthesis

The hypotheses for the synthesis method can be expressed as follows:

H_0: $\mu_T - \mu_P \leq \lambda_L \times (\mu_C - \mu_P)$ or $\mu_T - \mu_P \geq \lambda_U \times (\mu_C - \mu_P)$
versus
H_1: $\lambda_U \times (\mu_C - \mu_P) > \mu_T - \mu_P > \lambda_L\ (\mu_C - \mu_P)$

where λ_L and λ_U are the pre-specified preservation and inflation factors ($0 \leq \lambda_L \leq 1$ and $\lambda_U \geq 1$), respectively.[9] Equivalence is concluded if the test treatment preserves at least $100 \times \lambda_L\%$ and does not exceed $100 \times \lambda_U\%$ of the treatment effect observed with the active control in previous studies.

The synthesis method is designed to directly address the question of whether the test product would have been superior to a placebo had a placebo been in the study, and also to address the related question of what fraction of the active comparator's effect is maintained by the test product.[4,7] Use is made of the variability from the proposed trial and the historical trials and one confidence interval is constructed for testing the equivalence hypothesis that the treatment preserves a fixed fraction of the control effect, without actually specifying the size of the control effect (M1 in the fixed-margin approach) or a specific fixed equivalence margin based on the control effect. Clinical judgment is used to prespecify an acceptable fraction of the control therapy's effect (M2 in the fixed-margin approach) that should be retained by the test drug, regardless of the magnitude of the control effect.

The EMA View

The EMA published a guidance document on the choice of the Non inferiority (NI) margin in July 2005.[10] The guideline, which applies equally to the choice of equivalence margin, does not specify a method for choosing the margin. The choice of margin should be "based upon a combination of statistical reasoning and clinical judgement" and should provide evidence that the test product would have been shown to be efficacious if a placebo controlled trial had been performed.

The EMA argues that to adequately choose delta, an informed decision must be taken, supported by evidence of what is considered an unimportant difference in the particular disease area. If there are already many treatments being used interchangeably for the disease under consideration a possible approach might be to consider the information available from all of them. A delta may then be constructed based on the information known about the relative efficacy of these products. If there is only one product on the market then a "survey of practitioners on the range of differences that they consider to be unimportant" may be the best way to choose delta.

Choosing the Margin

All methods depend to some extent on historical data and clinical judgment and are inevitably subjective. Nonetheless, in order to ensure that decisions are as informed as possible, it is essential to gather all relevant information through a comprehensive literature search and/or evaluation of internal databases. The overall estimate of the treatment effect and population variability observed in historical studies with the active control should be determined quantitatively, ideally by meta-analysis.

We will now describe how an equivalence margin for biosimilar trastuzumab (HERCEPTIN) and biosimilar bevacizumab (AVASTIN) can be constructed using these techniques.

Trastuzumab

For the purpose of establishing the therapeutic equivalence of an oncology biosimilar, a randomized, double-blind, parallel-group, multicenter phase III trial comparing the efficacy, safety, immunogenicity, and pharmacokinetics in combination with backbone chemotherapy is usually performed. Trastuzumab is approved for several indications in Europe and the USA, including metastatic breast cancer (MBC), early breast cancer, and metastatic gastric cancer. The choice of target patient population is complex and involves balancing regulatory and scientific considerations with operational demands.[11] We will consider a study in patients with HER2-positive MBC.

The objective of a biosimilar study is not to demonstrate patient benefit per se, as that has already been demonstrated with the reference product, rather that the biosimilar medicine exhibits similar efficacy (and safety) to the reference medicinal product (the originator product). For this reason, the primary endpoint for biosimilar oncology studies would not be the traditional time-to-event endpoints such as overall survival (OS) or progression-free survival (PFS), but measures of activity, such as Overall Response Rate (ORR, proportion of patients in whom a Complete Response (CR) or Partial Response (PR) was observed). PFS and OS should also be included, but as part of the secondary endpoints.

So after having selected our target population and primary endpoint we can now focus on constructing the equivalence margin. The first step in establishing the equivalence margin is to identify all relevant publication pertaining to the proposed study. Table 1 outlines the relevant studies that were identified from

Table 1. Meta-analysis of Objective Response Rate (ORR) for Trastuzumab.

			Effect size	
Reference[a]/ Study number/name	Taxane + Trastuzumab	Taxane Only	Risk Difference	Risk Ratio
11, 15 / M77001 (with Docetaxel)	n = 92	n = 94		
	56 (60.9%)	34 (36.2%)	24.7%	1.67
11, 16 / H0648g[b] (with Paclitaxel)	n = 68	n = 77		
	33 (49%)	13 (17%)	31.7%	2.87
Combined effect size[c] (95% CI)			28.0%[d] (17.9%, 38.1%)	1.9[e] (1.452, 2.512)

[a] As per Reference list. Data taken from the "Herceptin: European Public Assessment Report—Scientific Discussion."[12]
[b] Subgroup of IHC3+ patients.
[c] Using the inverse variance-weighed method.
[d] *P* value of homogeneity test = .502.
[e] *P* value of homogeneity test = .095.

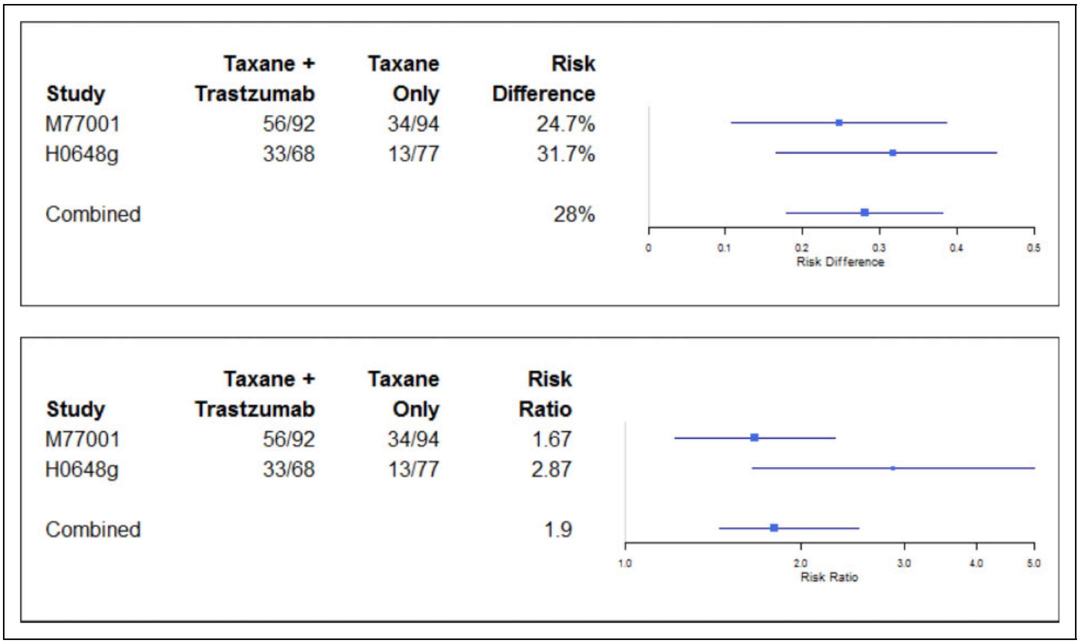

Figure 1. Meta-analysis of objective response rate (ORR) for trastuzumab.

1. Herceptin: European Public Assessment Report—Scientific Discussion[12]
2. Herceptin: European Public Assessment Report—Product Information[13]
3. FDA label for trastuzumab (Revised 02/2010)[14]
4. FDA Medical Statistical Review for trastuzumab[15]
5. Literature search for published randomized studies of trastuzumab for use in breast cancer using the key words: trastuzumab and breast cancer.

Two studies of trastuzumab with taxane in the treatment of patients with HER2-positive MBC were identified for the purposes of determining the effect size.[16,17] Table 1 and Figure 1 show the results of a meta-analysis of the proposed primary end point (ORR) from these studies. The combined effect size for the risk difference (28.0% [17.9%, 38.1%]; $P = .502$) and relative risk or risk ratio (1.9 [1.452, 2.512]; $P = .095$), which were calculated using the inverse variance-weighted method, was homogeneous across studies. The combined effect size or the lower bound of the 95% confidence interval of the risk difference (17.9%) or the relative risk (1.452) can be viewed as the largest acceptable margin (M1). From a clinical perspective, it is not uncommon to use 50% of M1 as the clinical margin (M2).[4,18]

It is noteworthy that the EMA and FDA approach to the use of risk difference and relative risk can vary, as evidenced by the recent approval of ABP 501, where the EMA stated that "It is considered acceptable to present the results as [risk ratio] RR," whereas the FDA stated that "a margin based on the absolute difference scale [should] be used, as it is considered more important than other metrics, such as risk ratio, from a clinical perspective for an evaluation of benefit-risk."[19,20] The use of

Table 2. M2 and Sample Sizes for Trastuzumab, Objective Response Rate, and Risk Difference.

Methodology	M2	Sample Size per Arm
Fixed-margin: 50% effect size	±14%	265[a]
Fixed-margin: 50% lower bound	±9%	640[a]
Synthesis	H_1: $p_e - p_c > -14\%$ and $p_e - p_c < 14\%$	203[b]

[a]Equivalence test for two proportions using differences, PASS 13 (80% power, actual difference = 0.0, alpha = 0.025, reference group proportion = 0.559).
[b]TOST (Two One-Sided Test) at the .05 level, 80% power, reference group proportion = 0.559.

Table 3. M2 and Sample Sizes for Trastuzumab, Objective Response Rate, Relative Risk.

Methodology	M2[a]	Sample Size per Arm[b]
Fixed-margin: 50% effect size	0.725, 1.378	167
Fixed-margin: 50% lower bound	0.830, 1.205	483

[a]Equivalence margin calculated as exp(ln[1/Δ]) × (1 − 0.5).
[b]Equivalence test for two proportions using ratios, PASS 13 (80% power, actual ratio =1.0, alpha = 0.025, reference group proportion = 0.559).

Table 4. M2 and Sample Sizes for Bevacizumab, Objective Response Rate, Risk Difference.

Methodology	M2[a]	Sample Size per Arm
Fixed-margin: 50% effect size	±9%	612[b]
Fixed-margin: 50% lower bound	±7%	1012[b]
Synthesis	H_1: $p_e - p_c > -9\%$ and $p_e - p_c < 9\%$	516[c]

[a]Risk difference from meta-analysis: 17.3% (13.7%, 20.9%); *P* value of homogeneity test = 0.1539.[22-24]
[b]Equivalence test for 2 proportions using differences, PASS 13 (80% power, actual difference = 0.0, alpha = 0.025, reference group proportion = 0.381).
[c]TOST (Two One-Sided Test) at the .05 level, 80% power, reference group proportion = 0.381.

Table 5. M2 and Sample Sizes for Bevacizumab, Objective Response Rate, Relative Risk.

Methodology	M2[a,b]	Sample Size per arm[c]
Fixed-margin: 50% effect size	0.735, 1.360	367
Fixed-margin: 50% lower bound	0.787, 1.271	601

[a]Relative risk from meta-analysis: 1.849 (1.615, 2.116); *P* value of homogeneity test = 0.224. Meta-analysis of previous studies with bevacizumab in patients with NSCLC.[22-24]
[b]Equivalence margin calculated as exp (ln[1/Δ]) × (1 − 0.5).
[c]Equivalence test for two proportions using ratios, PASS 13 (80% power, actual ratio =1.0, alpha = 0.025, reference group proportion = 0.381).

Table 6. M2 and Sample Sizes for Bevacizumab, Objective Response Rate, Risk Difference.

Methodology	M2[a]	Sample Size per Arm
Fixed-margin: 50% effect size	±10%	510[b]
Fixed-margin: 50% lower bound	±8%	800[b]
Synthesis	H_1: $p_e - p_c > -10\%$ and $p_e - p_c < 10\%$	423[c]

[a]Risk difference from meta-analysis: 20.6% (15.3%, 26.0%); *P* value of homogeneity test = 0.347. Meta-analysis of previous studies with bevacizumab in patients with NSCLC.[22,23,25]
[b]Equivalence test for 2 proportions using differences, PASS 13 (80% power, actual difference = 0.0, alpha = 0.025, reference group proportion = 0.402).
[c]TOST (Two One-Sided Test) at the .05 level, 80% power, reference group proportion = 0.381.

absolute difference or risk ratio also seems to depend on the study population. For example, we have seen the FDA request risk difference and relative risk for different indications for the same biosimilar compound.

Table 2 summarizes the sample sizes for each value of M2 for the fixed-margin and synthesis methods assuming a primary endpoint of Risk Difference. The synthesis margin was derived using "preservation" of effect method as discussed by Carrol and extended to biosimilar trials by Yining and Bin.[9,21] Table 3 presents analogous information assuming a primary endpoint of Relative Risk. There is as yet no published validated methodology for applying synthesis to equivalence studies with Relative Risk, but given that synthesis can be applied to both parameters in the non-inferiority setting, there is no reason why this cannot be done.[7] As can be seen, the method for determining the margin can significantly impact the number of patients to be recruited to demonstrate biosimilarity. The sample sizes required to establish equivalence range from 167 (50% of the relative risk) to 640 (fixed-margin using lower bound of the confidence interval) patients per group.

A like-for-like comparison of the synthesis and fixed-margin methods shows that the former is statistically more efficient (smaller sample size), as the standard error will always be smaller than that of the fixed-margin method.[4] There is as yet no published methodology for using the synthesis method in equivalence trials with risk ratio.

Bevacizumab

The above issues are not that dissimilar for other biosimilar compounds. Tables 4 and 5 show M2 and sample sizes for bevacizumab adopting the same strategy as above for trastuzumab with a primary endpoint of ORR, assuming a primary end point of Risk Difference and Relative Risk. The effect size for the risk difference and relative risk from a meta-analysis of studies in patients with non–small cell lung cancer (NSCLC) was 17.3% [13.7%, 20.9%] and 1.849 [1.615, 2.116], respectively.[22-24] The *P* values for homogeneity test were 0.1539 and 0.224, respectively. For bevacizumab, the range of sample sizes required to demonstrate therapeutic equivalence varies from 367 to 1012 patients per group with the synthesis method, again more statistically efficient than the fixed-margin

Table 7. M2 and Sample Sizes for Bevacizumab, Objective Response Rate, Relative Risk.

Methodology	M2[a,b]	Sample Size per Arm[c]
Fixed-margin: 50% effect size	0.679, 1.47	216
Fixed-margin: 50% lower bound	0.757, 1.320	411

[a]Relative risk from meta-analysis: 2.171 (1.743, 2.704); *P* value of homogeneity test = 0.683. Meta-analysis of previous studies with bevacizumab in patients with NSCLC.[22,23,25]
[b]Equivalence margin calculated as exp(ln[1/Δ]) × (1 – 0.5).
[c]Equivalence test for 2 proportions using ratios, PASS 13 (80% power, actual ratio =1.0, alpha = 0.025, reference group proportion = 0.402).

approach on a like-for-like comparison. Tables 6 and 7 show that the effect size for the risk difference and relative risk are highly dependent on the studies included in the meta-analysis.

Conclusion

The choice of acceptance margin for non-inferiority and equivalence studies is controversial.[4,18] As can be seen from the biosimilar trastuzumab (Herceptin) and biosimilar bevacizumab (Avastin) case studies, the resulting sample sizes differ significantly depending on the choice of margin and methodology used. Furthermore, the effect size on which the margin is decided can differ quite markedly depending on the studies included in the meta-analysis and the percentage of the treatment effect that is retained. In the recent approval of biosimilar Humira (adalimumab), the FDA questioned the studies included in the meta-analysis performed by the sponsor company.[26]

The synthesis method is more statistically efficient than the fixed-margin approach but, in our experience, is rarely if ever used in biosimilar studies. The reasons often given are the absence of a fixed margin, which renders the outcome of the study difficult to interpret, and the sensitivity of the method to the constancy assumption (ie, that the treatment effect observed in the equivalence study is consistent with that seen in historical trials). However, the margin is by its very nature highly subjective, and the outcome of such a study should not be the sole arbiter of whether a biosimilar product is equivalent (or not) to the originator. As stated in the regulatory guidance on biosimilars, the decision on biosimilarity should be based on the totality of the data and not on the outcome of any one study.[1,2] Furthermore, both the fixed-margin and the synthesis methods require the constancy assumption to be fulfilled.[4]

We argue that demonstrating that the biosimilar product retains a predefined fraction of the treatment effect should suffice with the proviso that comparability to the reference product in terms of physicochemical, in vitro functional characteristics, and nonclinical and clinical pharmacokinetic and pharmacodynamic profiles have all been demonstrated in well-designed studies.

Large comparative clinical trials defeat the object of the abbreviated development pathway for biosimilars adopted by regulatory authorities. Adoption of approaches such as the synthesis method for choosing the margin could potentially result in smaller studies and facilitate patient access to alternative therapy. We would therefore encourage biosimilar developers to discuss the use of this method with regulatory authorities during the planning stages of the pivotal phase 3 trial.

Author Note

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author).

Declaration of Conflicting Interests

No potential conflicts were declared.

Funding

No financial support of the research, authorship, and/or publication of this article was declared.

References

1. EMA. Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: nonclinical and clinical issues. EMEA/CHMP/BMWP/42832/2005 Rev. 1. July 1, 2015.
2. FDA. Guidance for industry. Scientific considerations in demonstrating biosimilarity to a reference product. *April* 2015.
3. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313:36-39.
4. Schumi J, Wittes J. Through the looking glass: understanding non-inferiority. *Trials*. 2011;12:106-117.
5. Rothmann M, Li N, Chen G, et al. Design and analysis of non-inferiority mortality trials in oncology. *Stat Med*. 2003; 22:239-264.
6. Rothmann MD, Tsou HH. On non-inferiority analysis based on delta-method confidence intervals. *J Biopharm Stat* 2003;13: 565-583.
7. FDA Draft Guidance for Industry, Non-Inferiority Clinical Trials, March 2010.
8. Combest AJ, Wang S, Healey BT, Reitsma DJ. Alternative statistical strategies for biosimilar drug development. *GaBI J*. 2014: 3:13-20.
9. Yining Y, Bin Y. Demonstrating biosimilarity via equivalence in clinical trials. *Stat Biopharm Res*. 2012;4:264-272.
10. CHMP. Guideline on the choice of non-inferiority margin, EMEA/CPMP/EWP/ 2158/99, *July* 2005.
11. Lai Z, La Noce A. Key design considerations on comparative clinical efficacy studies for biosimilars: adalimumab as an example. *RMD Open* 2016;2:e000154. doi:10.1136/rmdopen-2015000154.
12. Herceptin®: European public assessment report—scientific discussion. http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Scientific_Discussion/human/000278/WC500049816.pdf
13. Herceptin®: European public assessment report—product information 11/07/2011. Available at http://www.ema.europa.eu/docs/

en_GB/document_library/EPAR_-_Product_Information/human/000278/WC500074922.pdf

14. FDA label for Herceptin®. revised October 2010.
15. Herceptin®: FDA statistical review. http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/TherapeuticBiologicApplications/ucm091381.pdf.
16. Marty M, Cognetti F, Maraninchi D, et al. Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: The M77001 study group. *J Clin Oncol*. 2005;23(19): 4265-4274.
17. Slamon DJ, Leyland-Jones B, et al. Use of chemotherapy plus a monoclonal antibody against HER2 or metastatic breast cancer that overexpresses HER2. *N Engl J Med*. 2001; 344(11):783-792.
18. Wangge G, Roes KC, de Boer A, Hoes AW, Knol MJ. The challenges of determining noninferiority margins: a case study of noninferiority randomized controlled trials of novel oral anticoagulants. *CMAJ*. 2013;185(3):222-227.
19. Amgevita European Public Assessment Report (EPAR). EMA/106922/2017. January 26, 2017.
20. ABP501: Statistical review and evaluation. September 7, 2016.
21. Carroll KJ. Statistical issues and controversies in active-controlled, "noninferiority" trials. *Stat Biopharm Res*. 2013;5(3):229-238.
22. Sandler A, Gray R, Perry MC, Brahmer J, Schiller JH, et al. Paclitaxel carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med*. 2006;355:2542-2550.
23. Reck M, von Pawel J, Zatloukal P, Ramlau R, Gorbounova V, et al. Phase III trial of cisplatin plus gemcitabine with either placebo or bevacizumab as first line therapy for nonsquamous non-small-cell lung cancer: AVAil. *J Clin Oncol*. 2009;27:1227-1234.
24. Niho S1, Kunitoh H, Nokihara H, et al; JO19907 Study Group. 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.
25. 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: 2184-91.
26. FDA. Briefing document. Arthritis Advisory Committee meeting. BLA 761024, ABP 501, a proposed biosimilar to Humira (adalimumab). July 12, 2016.