The AAPS Journal (2019) 21: 47
DOI: 10.1208/s12248-019-0321-2

Research Article

Trial Design and Statistical Considerations on the Assessment of Pharmacodynamic Similarity

Peijuan Zhu,[1,4] Chyi-Hung Hsu,[1] Jason Liao,[2] Steven Xu,[1] Liping Zhang,[1] and Honghui Zhou[3]

Received 6 November 2018; accepted 17 March 2019; published online 3 April 2019

Abstract. Pharmacodynamics (PD) similarity is an important component to support the claim of similarity between two drugs or devices. This article investigates the trial design and statistical considerations in the equivalence test of PD endpoints. Using bone resorption marker CTX as a case study, the relationship between the PD readouts and drug potency was explored to evaluate the sensitivity of the PD endpoint and guide equivalence margin selection. For PD data that have high baseline variability, one conventional similarity assessment method was to apply baseline-normalization followed by the standard bioequivalence (BE) test (Lancet Haematol. 4:e350–61, 2017, Ann Rheum Dis. 2017). This study showcased the drawbacks of the conventional method for PD data that were close to inhibition saturation, as the baseline-normalization significantly skewed the distribution of the PD data toward non-log-normal. In such cases, the standard BE test can produce an inflated type I error. Alternatively, ANCOVA, when applied to the un-normalized PD data with the baseline as a covariate, produced a satisfactory type I error with sufficient power. Therefore, ANCOVA was recommended for equivalence test of PD markers that has a saturated inhibition profile and high variability at baseline. Moreover, the relationship between PD readouts and drug potency was used to explore the sensitivity of the PD endpoint and it could help justify the equivalence margins, since the standard 80% to 125% BE margin often does not apply to PD. Finally, a decision tree was proposed to help guide the design of the PD equivalence study in the choice of PD endpoints and statistical methods.

KEYWORDS: Analysis of covariance (ANCOVA); Equivalence; Baseline normalization; Pharmacodynamics (PD); Statistical test; Trial simulation.

INTRODUCTION

Pharmacodynamic (PD) similarity is an important component besides PK equivalence in the totality of evidence to support the claim of similarity between two drugs or devices. PD similarity can sometimes replace clinical efficacy comparison to support the demonstration of no clinically meaningful differences between a biosimilar and a reference product (3). It can also be applied to demonstrate the similarity of biologics which have undergone major post-approval manufacturing changes (4). Another important application is to support the claim of similarity when equivalence of PK cannot be demonstrated due to practical issues such as extreme low systemic exposure following local delivery (dermal, intravitreal, local gastrointestinal) or low bioavailability caused by high first-pass effect or poor absorption (5,6).

One of the major challenges to assessing PD similarity is that some PD endpoints are highly variable (typically greater than 50% coefficient of variation [CV]) (7–9) and therefore may require a large sample size. These PD endpoints are often baseline-normalized to reduce the inter-subject variability and hence reduce the sample size required (1,2,10). However, the baseline-normalization may skew the distribution of the PD data and render it non-log-normal. Since the standard BE test using geometric mean ratio (hereinafter referred to as "the standard BE test") assumes that the data follow a log-normal distribution (11), applying the standard BE test on baseline-normalized PD data may potentially lead to biased point estimates and may negatively impact the type I error and power of the study.

Another challenge in PD similarity assessment is the choice of proper margins for the equivalence test. The standard 80% to 125% margins typically used for PK BE assessment have been used to evaluate similarity on some PD endpoints (12–14). However, the standard 80% to 125%

Electronic supplementary material The online version of this article (https://doi.org/10.1208/s12248-019-0321-2) contains supplementary material, which is available to authorized users.

[1] Janssen Research and Development Inc, Raritan, NJ, USA.
[2] Merck & Co., Inc, North Wales, PA, USA.
[3] Janssen Research & Development Inc, Spring House, PA, USA.
[4] To whom correspondence should be addressed. (e–mail: pzhu9@its.jnj.com)

1550-7416/19/0300-0001/0 © 2019 American Association of Pharmaceutical Scientists

margins may not be appropriate for many PD endpoints due to their non-dose-proportional characteristics (3). Therefore, the choice of PD margins needs proper justification. When the correlation between PD and efficacy is well established, the PD margin may be derived based on clinically relevant efficacy margins, and this approach has been discussed extensively in the author's previous publications (3,15). When the correlation between PD and efficacy is not well characterized, the PD margin may be derived based on the relationship between PD and other key quality attributes such as drug potency (IC_{50}).

In this paper, the response of bone resorption biomarker, C-terminal telopeptide of type I collagen (CTX) after denosumab treatment was used as a case example to address the aforementioned challenges in demonstrating PD similarity. In osteoporosis, the CTX is a well-established bone resorption marker (16). Since denosumab inhibits receptor activator of nuclear factor kappa-B ligand (RANKL), a protein essential for the function of osteoclasts which are responsible for bone resorption, the CTX level decreases significantly after denosumab treatment (17,18). CTX is selected as the model PD endpoint because it represents a class of PD markers which follows a typical inhibition profile with a rapid onset, prolonged and near-complete inhibition followed by a slow recovery to baseline after drug washout. CTX is representative of many PD markers that share similar inhibition profiles, such as B cell depletion after rituximab treatment (1,19) and reduction of BCR-ABL transcript after imatinib treatment (20). In addition, the population PK/PD models for CTX with denosumab treatment are well published, making CTX a good candidate for the current simulation-based research.

Using published PK/PD models (7,17,18), the CTX profiles from a population treated by denosumab were simulated, and the area under the effect curve (AUEC) and the baseline-normalized AUEC (BN-AUEC) for CTX were derived. The relationships between drug potency IC_{50} and AUECs were explored. In addition, the distribution of BN-AUEC was assessed and, unlike that of AUEC, was found to be non-log-normal due to the presence of an upper bound corresponding to maximal (100%) inhibition. This observation triggered an investigation to evaluate the performance of the standard BE test regarding the type I error and power using trial simulations and to compare the standard BE test to other statistical methods such as bootstrap and ANCOVA to find an appropriate equivalence test method for similar PD endpoints. Finally, a decision tree was proposed to help guide the design of the PD equivalence study in the choice of PD endpoints and statistical methods based on the patient population, dose regimen, PD variability, and saturation.

METHODS

Bone Cycling Model Implementation and PD Endpoint Simulation

The bone resorption marker CTX model in response to denosumab treatment was implemented according to the published semi-mechanistic population PK/PD bone cycling model of CTX after denosumab treatment (7,17,21). The model was a closed form cyclical model with compartments representing sites of bone resorption, formation, primary mineralization, and secondary mineralization. Equations describing the time course of bone turnover biomarkers were developed using the flow rate of bone cycle units (BCU) between the compartments or the amount of BCU in each compartment with the incorporation of published bone turnover data. The treatment effect of denosumab was also incorporated as an inhibitory Emax model on the bone resorption rate parameters with denosumab concentration as an input. The model successfully described individual bone turnover markers including CTX as well as bone mineral density (BMD) results following treatment with denosumab in postmenopausal women (17), which made it suitable for clinical trial simulations of CTX data after denosumab treatment.

The above model was implemented in NONMEM (version 7.3; ICON development solutions, Ellicott City, MD). Since the clinical dose of denosumab in osteoporosis patients is 60 mg every 6 months (22), simulation of CTX single-dose profile after denosumab treatment was performed with a denosumab dose of 60 mg, and CTX readouts were simulated for the following timepoints to mimic a clinical trial sampling scheme: 0 week, then weekly from 0 to 16 weeks, then biweekly from 16 to 54 weeks.

AUEC Calculation

Baseline was defined as the CTX level immediately before the drug administration. Baseline-normalized CTX levels were calculated by division with baseline. The area under the effect curve from time 0 to t $AUEC_{(0\text{-}t)}$ and $BN\text{-}AUEC_{(0\text{-}t)}$ for CTX was calculated as the total area between the baseline and the CTX readout from time 0 to t. The CTX response had a pronounced rebound phenomenon after the denosumab washout, which was observed in the patient population, and was a characteristic incorporated in the model (17) (Fig. 1). Consequently, the proper calculation of AUEC required that the rebound, which introduced a partial negative area (area where CTX level was above the baseline), be addressed properly. For work described in this paper, $AUEC_{(0\text{-}t)}$ and $BN\text{-}AUEC_{(0\text{-}t)}$, with t equal to 3 months, 6 months, or 1 year, were calculated as the positive area between the baseline and the CTX readout. The negative area during the rebound was excluded from the $AUEC_{(0\text{-}t)}$ calculation to avoid the total AUEC being negative for some subjects. Another reason to exclude the rebound is that the rebound is a response inherent to the patients' physiology and is not directly driven by the drug effect, and therefore is less informative for PD similarity assessment. More importantly, in a real-world situation, the patients take repeated doses, and the rebound effect is not present until the end of treatment.

The AUEC calculation was implemented by inserting a PD readout of 0 at time $t_{cross\text{-}bl}$ (the time that CTX crossed baseline) using the time and CTX level immediate before (t_{below_bl}, $CTX_{below\text{-}bl}$) and after (t_{above_bl}, $CTX_{above\text{-}bl}$) crossing the baseline of CTX (CTX_{bl}), assuming linear interpolation according to Eq. 1. The $AUEC_{(0\text{-}t)}$ was then calculated using the trapezoidal rule in R 3.4 (CRAN package PK) where t was equal to $t_{cross\text{-}bl}$ or the predefined t, whichever was earlier.

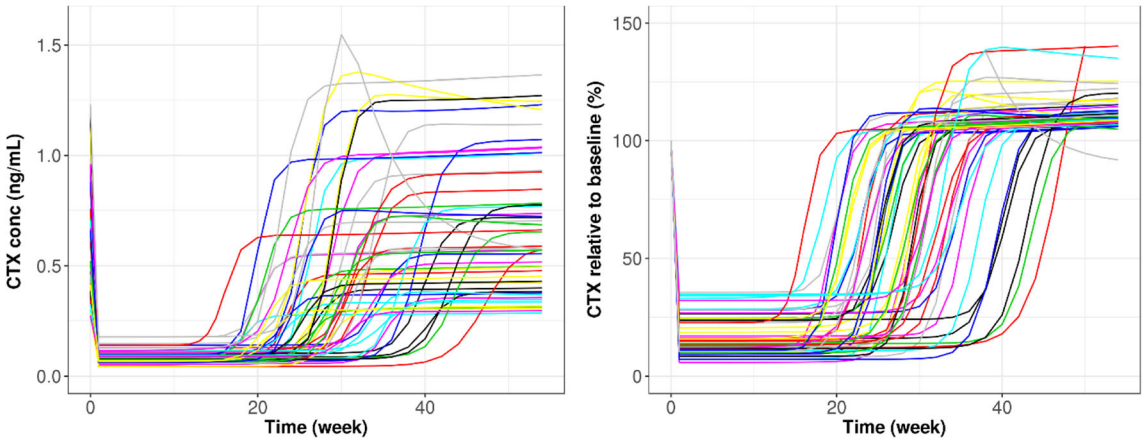

Fig. 1. Simulated CTX profile (left) and baseline-normalized CTX profile (right)

$$t_{cross-bl} = t_{below\_bl} + \frac{CTX_{bl} - CTX_{below\_bl}}{CTX_{above\_bl} - CTX_{below\_bl}} \times (t_{above_{bl}} - t_{below_{bl}}) \quad (1)$$

The log-normal distribution assumptions of AUECs and BN-AUECs were evaluated using histograms and Q-Q plots.

The AUEC versus potency relationship was illustrated by simulating the CTX profiles using the above-described method, assuming relative IC_{50} values at 0.1, 0.2, 0.4, 1, 2, 4, and 10 times the IC_{50} of denosumab, with all other PK and PD parameters identical to those of denosumab. The population median $AUEC_{(0\text{-}t)}$ of CTX (AUEC) or baseline-normalized CTX (BN_AUEC) was then plotted against the relative IC_{50}.

Trial Simulations

Assuming that a presumable biosimilar is being compared with the reference drug denosumab using CTX AUEC as the PD equivalence endpoint, a set of equivalence margins need to be selected. The equivalence margin defines a range of values for which the efficacies are "close enough" to be considered equivalent. In practical terms, the margin is the maximum clinically acceptable difference that one is willing to accept in return for the benefits of a new therapy (23). For the standard BE test on the $BN\_AUEC_{(0\text{-}6m)}$ data (see "DISCUSSION" section for the reason of choosing $AUEC_{(0\text{-}6m)}$), an arbitrary symmetrical equivalence margin was selected to detect a 10-fold or higher difference in IC_{50} (10-fold lower potency) based on the AUEC-potency relationship (see "RESULTS" section), as one may be concerned with a biosimilar product that has low potency. According to the choice of margin based on the potency, three patient populations each composed of 50,000 subjects were simulated. The reference population was simulated assuming a single 60-mg dose of denosumab. The test population 1 was simulated assuming a single 60-mg dose of a presumable denosumab biosimilar that has an IC_{50} at 10-fold that of denosumab and with all other PK/PD parameters being identical to denosumab. Similarly, the test population 2 was simulated assuming a single 60-mg dose of a presumable denosumab biosimilar that has an IC_{50} at 1/10 that of denosumab.

Consequently, the population median $BN\_AUEC_{(0\text{-}6m)}$ of test population 1 was set as the lower margin (0.96). The upper equivalence margin was set to 1.04 (1/0.96) based on the symmetry to the lower margin in the log-transformed space, due to the requirement of symmetrical margins for a standard BE test. To create a population with a median $BN\_AUEC_{(0\text{-}6m)}$ at the upper margin, the test population 2 was shifted in the log-transformed space then reversed back to the original scale.

Trial simulations were set up to compare the presumable biosimilar with the reference drug denosumab. The trial sample size of 124 per arm was calculated using the two one-sided test method for the standard BE test design (24) assuming the above-mentioned margins (0.96, 1.04) for $BN\_AUEC_{(0\text{-}6m)}$, two-sided type I error (alpha) of 10%, test to reference mean ratio of 1, reference CV of 10%, and 90% power. In scenario 1, it was assumed that test population 1 was the test group and the reference population was the reference group. The probability of falsely rejecting the null hypothesis (type I error) when the test group was at the lower margin was assessed. In scenario 2, it was assumed that test population 2 was the test group and the reference population was the reference group. The probability of falsely rejecting the null hypothesis (type I error) when the test group was at the upper margin was assessed. In scenario 3, it was assumed that the test group and the reference group were identical to the reference population. The probability of correctly rejecting the null hypothesis (power) when the test group was identical to the reference group was assessed.

The trial simulations and equivalence tests were implemented in R (3.4).

Standard BE Test

For each trial of 124 subjects per arm, the standard BE test, i.e., the *two* one-*sided* test, using 90% confidence interval (CI) of test-to-reference geometric-mean-ratio was performed on $BN\_AUEC_{(0\text{-}6m)}$ data with the null hypothesis that the population mean of log-transformed $BN\_AUEC_{(0\text{-}6m)}$ was different between test and reference. The type I error was calculated for scenario 1 and scenario 2 as the percentage of trials with 90% CI of geometric-mean-ratio falling within the

margin (0.96, 1.04). Power was calculated for scenario 3 as the percentage of trials with 90% CI of geometric-mean-ratio falling within the margin.

Bootstrap

Bootstrap was performed on the same set of trial simulation data as that used in the standard BE test. For each trial of 124 subjects per arm, the 90% CI of the median difference of log-transformed $BN\_AUEC_{(0\text{-}6m)}$ was calculated using the quantiles of 5% and 95% from the bootstrap method by sampling 5000 times with replacement to generate a trial dataset with 124 subjects per arm from each simulated trial. Then, it was determined whether the 90% CI of the bootstrapped median ratio (the reverse log-transformed median difference) fell within the margin (0.96, 1.04). The method was evaluated on the same trial simulation data of $BN\_AUEC_{(0\text{-}6m)}$ as described above for the standard BE test and its performance was compared to the standard BE test.

ANCOVA

ANCOVA was performed on the same set of trial simulation data as that used in the standard BE test. Correlation between AUEC and baseline was evaluated by plotting $\log(AUEC_{(0\text{-}6m)})$ versus $\log(CTX_{bl})$. It was revealed that most of the variability of the $AUEC_{(0\text{-}6m)}$ could be explained by the variability in the baseline (see "RESULTS"). This observation suggested that ANCOVA may be a good candidate for the equivalence test of $CTX\_AUEC_{(0\text{-}6m)}$ with CTX_{bl} incorporated as the covariate. This approach may address the high variability problem without baseline-normalization so that the log-normality holds true. Assuming the same trial design as described above for the standard BE test, 5000 trial simulations (124 subjects per arm) were performed for scenarios 1, 2, and 3 and ANCOVA was applied to the $AUEC_{(0\text{-}6m)}$ data with CTX_{bl} as the covariate. For subject i, the regression is outlined in Eq. 2 with treatment being 0 for the reference and 1 for the test. The $\exp(\beta 2)$ is the ratio of $AUEC_{(0\text{-}6m)}$ between the test and reference group. The margins for $AUEC_{(0\text{-}6m)}$ were 0.96 and 1.04, which was calculated in the same fashion as that for $BN\_AUEC_{(0\text{-}6m)}$, by taking the median $AUEC_{(0\text{-}6m)}$ of test population 1 as the lower margin. The upper margin was calculated symmetrically in the log-transformed space. The margins were almost identical between $AUEC_{(0\text{-}6m)}$ and $BN\_AUEC_{(0\text{-}6m)}$, which demonstrated that the AUEC and baseline-normalized AUEC had similar relative changes with regard to potency (see "RESULTS"). The 90% CI of $\exp(\beta 2)$ was compared to the margin of (0.96, 1.04) to determine whether the equivalence is met. The type I error and power of the ANCOVA method were compared to that of the standard BE test and the bootstrap method.

$$\log\left(CTXAUEC_{(0\text{–}6m)_i}\right) = \beta 0 + \beta 1 \times \log(CTX_{bli}) + \beta 2 \times treatment_i + \epsilon_i \quad (2)$$

Parameter Sensitivity Analysis

Three additional sets of trial simulations were conducted by incorporating the parameter uncertainty in the PK and PD models as described in the publications (7,17,21) by randomly generating each parameter based on a normal distribution with the mean and standard error of the estimates. The standard BE test and ANCOVA methods were tested in these three sets of trial simulations using the same method as described above and the results were compared with the primary analysis

RESULTS

CTX Profiles

Using published population PK/PD bone cycling model which described individual bone turnover marker CTX results following denosumab treatment in postmenopausal women (7,17,18), the CTX profiles were simulated, and representative profiles of randomly selected 80 subjects are shown in Fig. 1. After denosumab treatment, CTX level dropped rapidly, reaching maximal inhibition after 1 week. The maximal inhibition was sustained for around 20 weeks for most subjects, then slowly returned to baseline after denosumab washout. For some subjects, a rebound was seen when the CTX level went above baseline then slowly regressed to baseline. This rebound effect was observed in clinical subjects (17), and the model was able to capture this characteristic.

Distribution of AUEC Data for CTX

Based on the simulated CTX data after a single dose of denosumab, $AUEC_{(0\text{-}3m)}$, $AUEC_{(0\text{-}6m)}$, and $AUEC_{(0\text{-}1y)}$ (AUEC from time 0 to t, with t at 3 months, 6 months, or 1 year) and the corresponding baseline-normalized AUEC data were calculated, respectively. The distributions of AUEC and BN_AUEC were explored using histograms and Q-Q plots. Histograms and Q-Q plots of $AUEC_{(0\text{-}6m)}$ and $BN\_AUEC_{(0\text{-}6m)}$ are shown in Figs. 2 and 4, respectively. Histograms comparing BN-AUEC at 3 months, 6 months, and 1 year are shown in Fig. 3. Although the distribution of the CTX AUEC data was log-normal, the distributions of the $BN\_AUEC_{(0\text{-}t)}$ data with a relatively short t, e.g., $BN\_AUEC_{(0\text{-}3m)}$ and $BN\_AUEC_{(0\text{-}6m)}$, were non-log-normal. Their distributions had a heavy left tail and an upper bound on the right, which was due to the fact that the majority of subjects had CTX level close to saturation or maximal inhibition between time 0 and t (t being 3 months or 6 months) (Fig. 2). In addition, the Q-Q plots showed that log-transformation on $BN\_AUEC_{(0\text{-}3m)}$ and $BN\_AUEC_{(0\text{-}6m)}$ did not bring the distribution close to normal, but instead further skewed the distribution away from normal (Figs. 2 and 4). These results suggest that the standard BE test, which is based on a log-normal-distribution assumption, may introduce biased results when applied to the baseline-normalized endpoints like $BN\_AUEC_{(0\text{-}3m)}$ and $BN\_AUEC_{(0\text{-}6m)}$. Interestingly, the distribution of $BN\_AUEC_{(0\text{-}1y)}$ was close to log-normal (Fig. 3), which may be because that the CTX level in most patients returned to baseline between weeks 20 and 40 and the average inhibition of CTX from time 0 to 1 year is far from saturation. As a consequence, the distribution of $AUEC_{(0\text{-}1y)}$ was not significantly affected by the upper bound and was close to log-normal. Therefore, the standard BE test would be suitable for $BN\_AUEC_{(0\text{-}1y)}$.

As expected, the baseline normalization significantly reduced the variability of the AUEC data. The CV for $CTX\_AUEC_{(0\text{-}6m)}$ was 48% while that for CTX_BN_

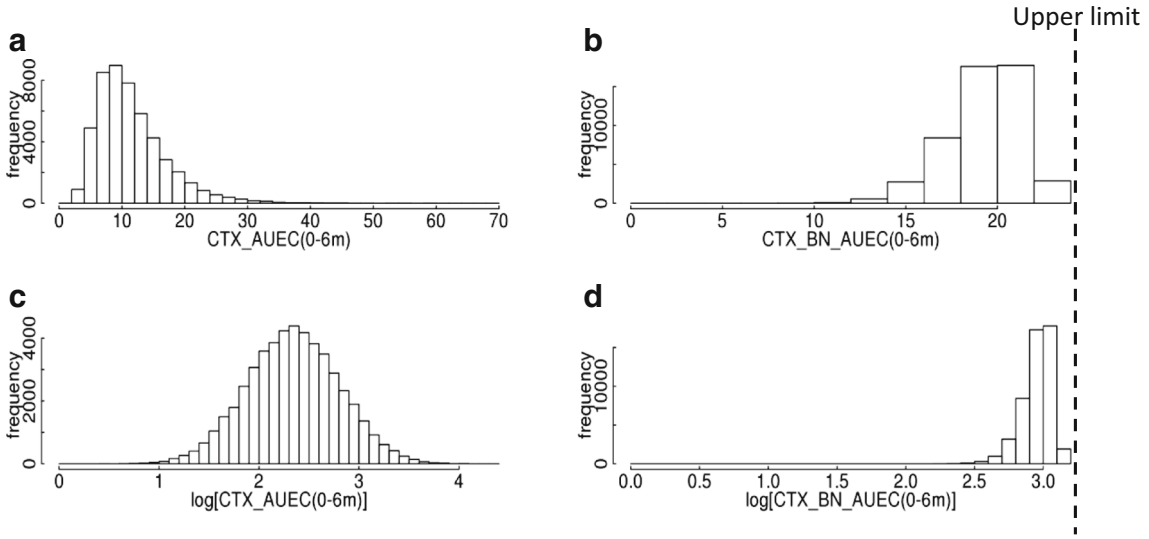

Fig. 2. Distribution histogram of CTX $AUEC_{(0\text{-}6m)}$ and baseline-normalized CTX $AUEC_{(0\text{-}6m)}$. **a** CTX $AUEC_{(0\text{-}6m)}$; **b** baseline-normalized CTX $AUEC_{(0\text{-}6m)}$; **c** CTX $AUEC_{(0\text{-}6m)}$ after log-transformation; **d** baseline-normalized CTX $AUEC_{(0\text{-}6m)}$ after log-transformation. The dotted line indicates the upper limit which was imposed by the maximal inhibition of 100%

$AUEC_{(0\text{-}6m)}$ was only 10%. Similar reductions were observed for AUECs at 3 months. The CV for CTX_ $AUEC_{(0\text{-}1y)}$ was 55% while that for CTX_BN_ $AUEC_{(0\text{-}1y)}$ was only 26%.

AUEC Versus Potency

We further explored the AUEC and potency (IC_{50}) relationship which is shown in Fig. 5. As the IC_{50} increased or the potency decreased, the $AUEC_{(0\text{-}6m)}$ decreased accordingly. However, the magnitude of the decrease was low relative to the decrease in potency. For a 10-fold increase in IC_{50} (10-fold decrease in potency), the decrease in both $AUEC_{(0\text{-}6m)}$ and $BN\_AUEC_{(0\text{-}6m)}$ was about 4%. Moreover, for a 10-fold decrease in IC_{50} (10-fold increase in potency), the increase in both $AUEC_{(0\text{-}6m)}$ and $BN\_AUEC_{(0\text{-}6m)}$ was merely 1%. This was because the clinical dose of 60 mg every 6 months was already a saturating dose, and a further increase in potency would not significantly increase the $AUEC_{(0\text{-}6m)}$.

Standard BE Test

If the PD equivalence of CTX is to be demonstrated in patients with a dosing schedule of 60 mg per 6 months per the US label, $BN\_AUEC_{(0\text{-}6m)}$ would most likely be chosen as the PD endpoints as it assesses the average PD response during the entirety of the first dosing cycle (see "DISCUSSION"). Since the distribution of $BN\_AUEC_{(0\text{-}6m)}$ was non-log-normal, clinical trial simulations were

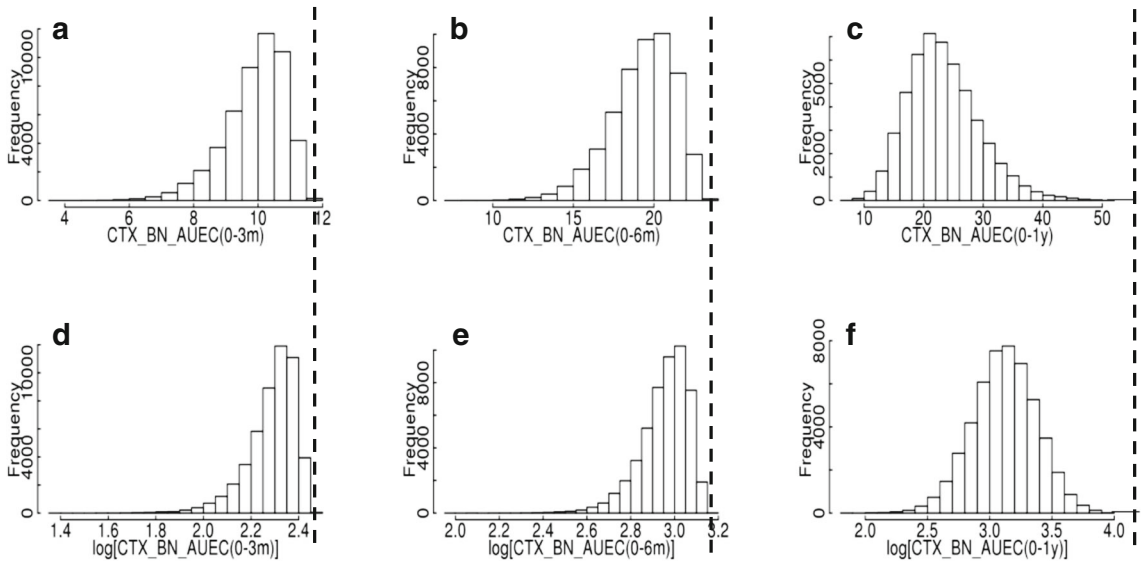

Fig. 3. Distribution histograms of baseline-normalized CTX AUEC at 3 months, 6 months and 1 year. The dotted line indicates the upper limit which was imposed by the maximal inhibition of 100%

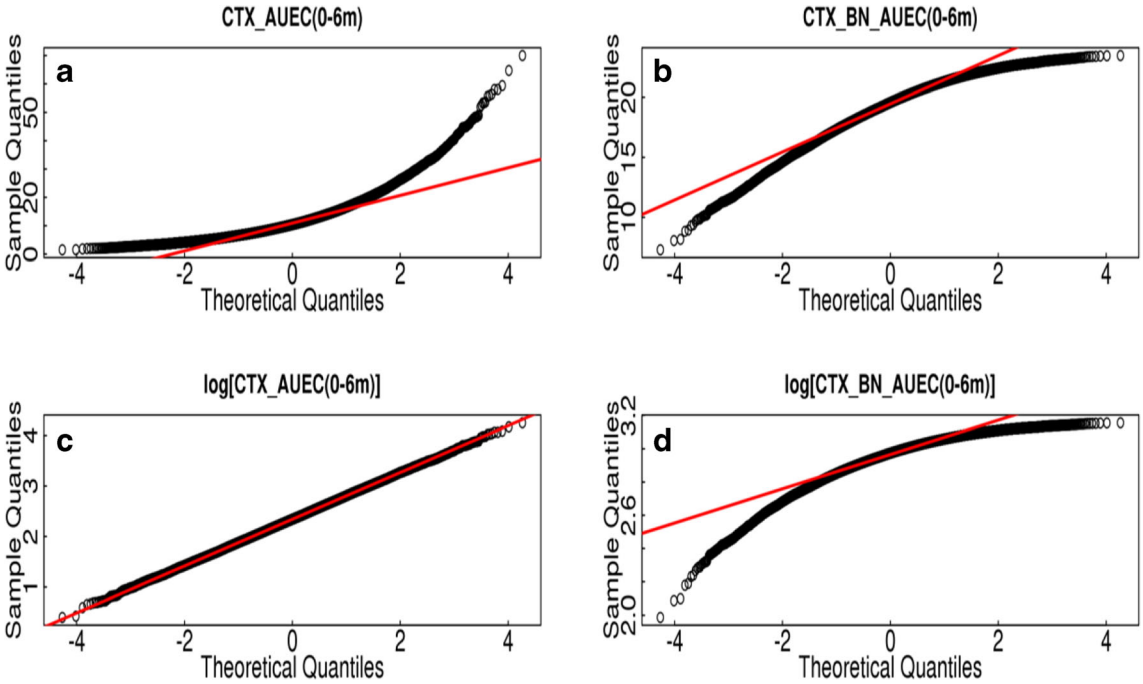

Fig. 4. Q-Q plot of CTX $AUEC_{(0\text{-}6m)}$ and baseline-normalized CTX $AUEC_{(0\text{-}6m)}$. **a** CTX $AUEC_{(0\text{-}6m)}$; **b** baseline-normalized CTX $AUEC_{(0\text{-}6m)}$; **c** CTX $AUEC_{(0\text{-}6m)}$ after log-transformation; **d** baseline-normalized CTX $AUEC_{(0\text{-}6m)}$ after log-transformation

performed to evaluate the performance of the standard BE test on BN_$AUEC_{(0\text{-}6m)}$ regarding the type I error and power. To detect a 10-fold difference in potency relative to the reference, the arbitrary lower margin was set at 0.96, the relative BN_$AUEC_{(0\text{-}6m)}$ value when the IC_{50} is 10-fold that of denosumab. The reason that the arbitrary symmetrical margins were benchmarked on the higher bound of 10-fold IC_{50} in this case study was that a lower potency (or higher IC_{50}) would be a major concern for a presumable biosimilar. The upper margin was set at 1.04 based on symmetry in the log space since the standard BE test requires symmetrical margins.

Five thousand (5000) clinical trials were simulated for three scenarios to evaluate the type I error and power of the standard BE test (see "METHODS"). The performance metrics of the standard BE test on BN_$AUEC_{(0\text{-}6m)}$ are shown in Table I. Because the distribution of BN_$AUEC_{(0\text{-}6m)}$ was non-log-normal, the type I error was higher than the designed 10% in both scenarios 1 and 2, with the test group on the lower or higher margin, respectively. The bias of the point estimate was 0.323% and −0.150% in scenarios 1 and 2, respectively. In addition, the power of the standard BE test assessed in scenario 3 with the test group being identical to the reference group was 83.8%, similar to the designed 90% power.

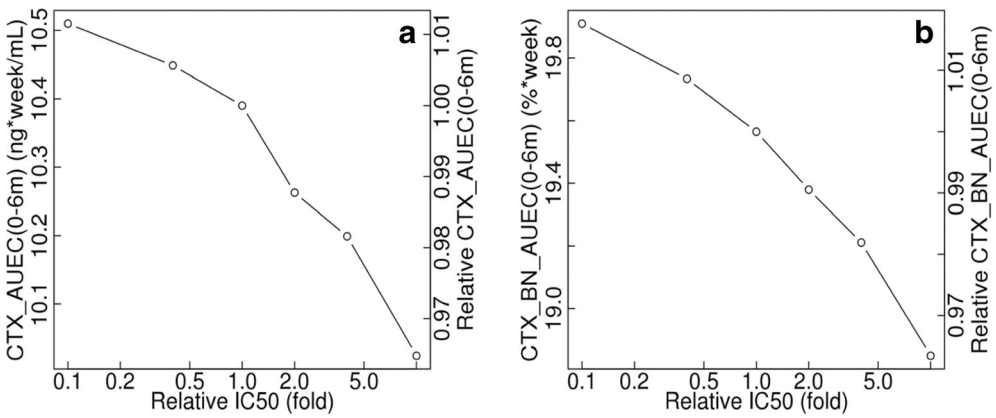

Fig. 5. The relationship between CTX AUEC and potency (IC_{50}). **a** population median CTX_ $AUEC_{(0\text{-}6m)}$ versus relative IC_{50} in comparison to the IC_{50} of denosumab; **b** population median CTX_ BN_$AUEC_{(0\text{-}6m)}$ versus relative IC_{50} in comparison to the IC_{50} of denosumab. The *Y*-axis on the left side indicates absolute values of the $AUEC_{(0\text{-}6m)}$ and CTX_ BN_$AUEC_{(0\text{-}6m)}$. The Y-axis on the right indicates relative values of population median $AUEC_{(0\text{-}6m)}$ and population median CTX_ BN_$AUEC_{(0\text{-}6m)}$ compared to those of denosumab

Table I. Comparison of Model Performance: Standard BE Test, Bootstrap, and ANCOVA

Parameter	Test population 1 (lower margin) versus reference			Test population 2 (upper margin) versus reference			Test population (identical to reference) vs reference		
	Standard BE	Bootstrap	ANCOVA	Standard BE	Bootstrap	ANCOVA	Standard BE	Bootstrap	ANCOVA
Type I error (design at 10%)	14.1%	5.18%	7.14%	12.3%	6.60%	1.94%			
Power (design at 90%)							83.8%	38.2%	86.9%
CI width (relative to reference)	3.60%	5.61%	3.84%	3.66%	5.57%	3.77%	3.58%	5.59%	3.73%
Bias of point estimate of difference (relative to reference)	0.323%	−5.57e-3%	0.214%	−0.150%	−7.45e-3%	0.512%	−0.021%	5.55e-3%	−7.57e-4%

Bootstrap

The bootstrap method was evaluated with the same trial simulation data used in the standard BE test, and the performance metrics of the bootstrap method are shown in Table I. The bootstrap method produced type I errors of 5.18 and 6.60% on the lower and upper margins, respectively, as well as low power (38.2%). These results seem to be due to a wider 90% CI produced with the bootstrap method at 5.6% in comparison to the 90% CI of 3.6% for the standard BE test. The wider CI may be due to the large variability from sampling a small sample (124 subjects per arm) with replacement. The bias of point estimate was minimal for the bootstrap method on the robust median metric, which was expected since bootstrap is a non-parametric method and does not rely on any assumptions of underlying distributions.

ANCOVA

Due to the high correlation observed between AUEC(0-6m) and baseline CTX levels (Fig. 6), the ANCOVA method was applied to the un-normalized $AUEC_{(0\text{-}6m)}$ data from the same trial simulation dataset used for standard BE test, with CTX_{bl} as a covariate. The performance metrics of the ANCOVA method are shown in Table I. The ANCOVA method produced type I errors well below 10% for scenarios 1 and 2. The bias of the point estimate was 0.214%, lower than the 0.323% of standard BE test on the lower margin, while on the upper margin, the bias was 0.512%. Although the magnitude of the bias was large on the upper margin, it was trending toward over-predicting the difference and resulted in a very low type I error. Moreover, the power of the ANCOVA method was at 86.9%, similar to the designed 90% power. The width of the 90% CI of the ANCOVA method was around 3.8%, similar to that of the standard BE method at around 3.6%.

Parameter Sensitivity Analysis

The standard BE test and ANCOVA methods were compared again in the parameter sensitivity analysis, in which the parameter uncertainty in the PK and PD models was incorporated by randomly generating three sets of parameters based on a normal distribution with the mean and standard error of the estimates. The results of parameter sensitivity analysis agreed closely with the primary analysis results shown in Table I. This was expected as the parameter uncertainty of the PK and PD model is low relatively to the corresponding between-subject variability (7,17,21) and the trial simulation results are highly consistent.

DISCUSSIONS

The sensitivity is always an important factor to consider when selecting a clinical endpoint for the equivalence test. It has long been realized that PD endpoints are often more sensitive than clinical endpoints (3). As outlined in the FDA's clinical pharmacology guidance for biosimilar development, the PK/PD study can potentially replace clinical efficacy comparison for the demonstration of biosimilarity (25). Although more sensitive than clinical efficacy, the clinical PD endpoints may still have limited sensitivity in comparison to *in vitro* potency assays, which usually can easily detect a 20–50% difference in potency (26). As shown in the CTX example, a 10-fold increase in IC_{50} (10-fold lower potency) only led to ~4% reduction in median $AUEC_{(0\text{-}6m)}$. Health authorities are aware of this issue, and the choice of PD margins often requires justification from the sponsors on a case by case basis. Hence, no uniform equivalence margins are likely to exist for PD endpoints as in the case of PK. In this paper, we had chosen an arbitrary margin for AUEC endpoints which corresponded to a 10-fold difference in potency to enable the statistical analysis. This 10-fold margin chosen in this manuscript does not represent a recommendation from the authors and any choice of margins should have its scientific rationale and need to be discussed with health authorities. Besides potency, margin selection can also take into consideration the correlation between PD endpoints and clinical efficacy if the relationship is known (3,15). A carefully selected margin may require a thorough understanding of the interplay between the PD response and drug exposure, drug potency, drug efficacy, etc.

AUEC is often the endpoint for PD markers as it reflects the average effect over the entire treatment course between time 0 to t, rather than an effect observed at a single time point. AUEC is listed as one of the primary choices of PD endpoints in the FDA's

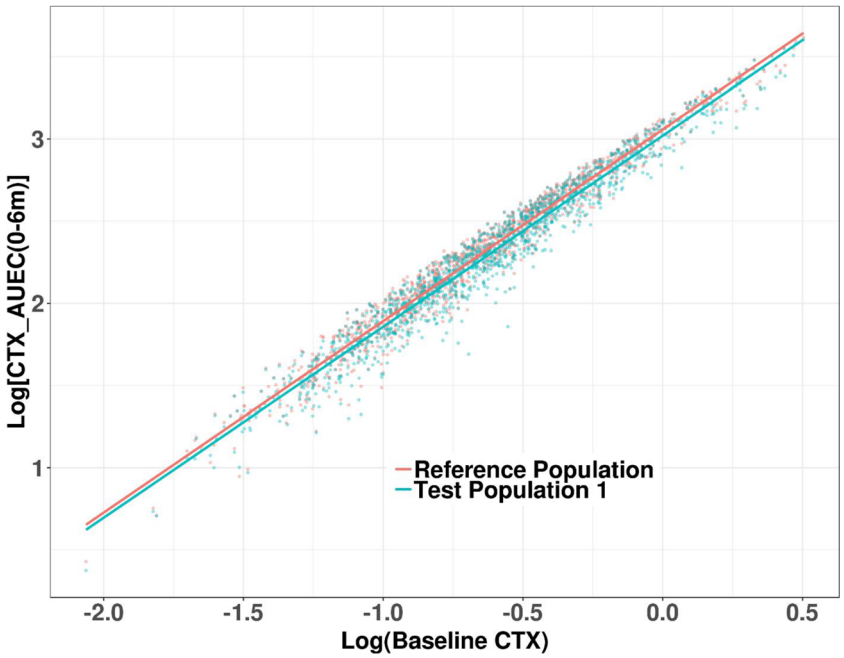

Fig. 6. High correlation between $AUEC_{(0\text{-}6m)}$ and baseline CTX levels

clinical pharmacology guidance on biosimilar development (25). If the PD similarity of CTX between a proposed biosimilar and the reference product is to be demonstrated in a patient study with a dose of 60 mg every 6 months per the product label, $AUEC_{(0\text{-}6m)}$ or $BN\_AUEC_{(0\text{-}6m)}$ may be chosen as the PD endpoints for equivalence test as it assesses the average PD response during the entirety of the first dosing cycle. AUEC measurements at steady state may not be the best PD endpoints to evaluate equivalence due to the lack of a well-defined immediately preceding baseline, although descriptive PD data at steady state may be assessed as secondary PD endpoints. Empirically, descriptive analysis of partial $AUEC_{(0\text{-}t)}$ with t less than 6 months (e.g., $AUEC_{(0\text{-}1m)}$ and $AUEC_{(0\text{-}3m)}$) is often requested by the health authorities to add to the totality of evidence for biosimilarity.

Whether to choose healthy subjects or patients for PD similarity assessment is mainly determined based on whether the PD responses are similar between the two populations. There are other considerations such as whether safety will be incorporated in the study and these considerations have been elaborated in previous publications (3,15). In cases where the primary PD similarity can be established in healthy subjects, a single-dose study design may be chosen. In that case, the most appropriate PD endpoint for equivalence test may be $AUEC_{(0\text{-}1y)}$ or $BN\_AUEC_{(0\text{-}1y)}$ as it captures the entire CTX profile from the onset of rapid inhibition to the complete recovery to baseline. Furthermore, the distribution of $BN\_AUEC_{(0\text{-}1y)}$ is close to log-normal (Fig. 2) and has a much lower variability of 26% CV in comparison to the 55%

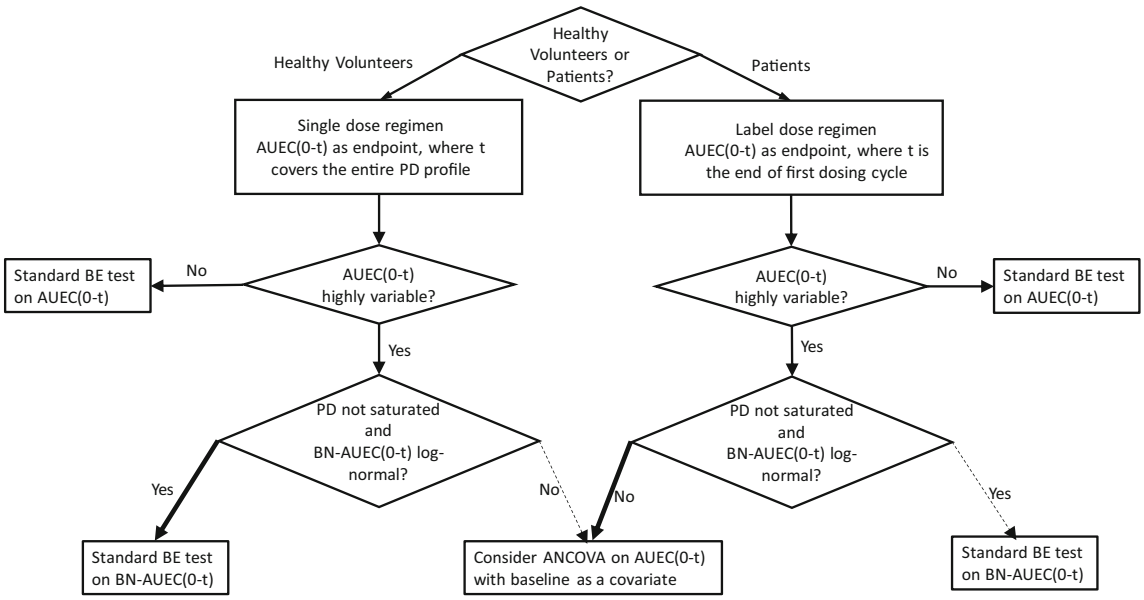

Fig. 7. Proposed decision tree for the choice of PD endpoint and equivalence test method

CV for $AUEC_{(0\text{-}1y)}$. The $BN\_AUEC_{(0\text{-}1y)}$ is, therefore, suitable to be assessed using the standard BE test. In this case, the standard BE test would be the preferred method due to the regulatory preference for methods that are simple and have few model-based assumptions.

In the case of a patient study, AUECs during the first dosing interval are often the preferred PD endpoints, i.e., $BN\_AUEC_{(0\text{-}6m)}$ or $AUEC_{(0\text{-}6m)}$. Although $BN\_AUEC_{(0\text{-}6m)}$ is significantly less variable with 10% CV in comparison to the 48% CV for $AUEC_{(0\text{-}6m)}$, the use of standard BE test on $BN\_AUEC_{(0\text{-}6m)}$ may not be appropriate if the PD is close to saturation and the distribution of PD data is non-log-normal, which could lead to inflated type I error as shown in the case of CTX. Alternative methods were explored in this study. The Bootstrap method, being a non-parametric method, was able to maintain a low type I error rate but suffered a problem of low power due to a wider 90% CI. The ANCOVA method of equivalence test on $AUEC_{(0\text{-}6m)}$ with CTX_{bl} as a covariate was shown to be the best method as the type I error was maintained below 10% and the power was preserved at a level close to 90%.

The better performance of ANCOVA in this case study was not surprising, as it was conducted on the un-normalized AUEC data, which followed a log-normal distribution. Thus, the underlying distribution assumption was not violated. Moreover, ANCOVA can help remove the impact of one or more independent variables (covariates) from the dependent variable before comparing the treatment effects. With the incorporation of baseline as a covariate, which accounted for most of the variability in the PD data and consequently reduced the variability of the PD response due to drug treatments, the ANCOVA method made it possible to establish equivalence with a relatively small sample size.

The preference of ANCOVA over baseline normalization has been discussed in a few publications. One paper published by Vickers in 2001 performed evaluation of superiority test on pain data simulated from a simple normal distribution (27). Four analysis endpoints were compared: post-treatment readout; absolute change between baseline and post-treatment; change normalized by baseline and ANCOVA of un-normalized post-treatment readout with the baseline as a covariate. The author concluded that for superiority test of a continuous outcome with both baseline and post-treatment readouts, ANCOVA has the highest statistical power and the change normalized by baseline has the lowest statistical power and was highly sensitive to changes in variance (27). The type I error was not evaluated in this publication. In another essay published by Zhang *et al.* (6), it was demonstrated by simulation that for superiority test, change normalized by baseline had sometimes better and other times worse statistical power in comparison to the absolute change between baseline and post-treatment, depending on whether the simulation of changes was done by assuming fixed absolute change or fixed percentage change (28).

The above results from Zhang's essay highlighted the limitation of using randomly simulated data from a normal distribution to evaluate statistical method performance since the arbitrary simulation setup has a strong impact on the statistical test results. The simulated data may be different from clinical data with regard to distributions, and the statistical test would perform differently with real clinical data in comparison to using randomly simulated data. Moreover, Vickers's publication did acknowledge that the baseline normalization may make the data non-normally distributed (27) but did not investigate in detail how the baseline normalization may affect the distribution of the data and how it may impact the bias of point estimate and therefore the type I error. In our current paper, to ensure that the data used for statistical method evaluation are representative of real clinical data, we used the semi-mechanistic population PK/PD bone cycling model of CTX to simulation patient CTX data after denosumab treatment. Since the model successfully described individual CTX following treatment with denosumab in postmenopausal women (17), the trial simulation data described in our current paper were representative of clinical data, and thus provided exceptional value in assessing the performance of different statistical methods using close-to-real datasets. In addition, most publications focused on superiority trials, and relatively few publications evaluated statistical methods for equivalence trials. In addition, the application of ANCOVA in equivalence test has mostly been limited to PK (29,30) and there were few reports on PD. This paper, therefore, provides unique values by filling in these gaps.

CONCLUSIONS

In this paper, a case study was used to demonstrate how an equivalence margin and an appropriate statistical method can be chosen to determine the PD biosimilarity.

The PD versus potency relationship can be helpful in selecting margins for PD equivalence test. PD marker such as CTX, although much more sensitive than the clinical efficacy endpoint such as bone mineral density (BMD) or frequency of bone fractures, still has limited sensitivity in comparison to *in vitro* potency assays. The traditional 0.80 to 1.25 PK equivalence margin may not necessarily be applicable to many PD endpoints, and sponsors would likely be requested by the health authorities to justify their choice of PD equivalence margins based on their knowledge of correlations between the PD response and drug exposure, drug potency, drug efficacy, etc.

The CTX case study in our current study showed that baseline normalization, although significantly reduced inter-subject variability, had a potential to alter the distribution of the PD data and make it non-log-normal, and therefore not suitable for the standard BE test. This phenomenon would become prominent when the PD readouts were close to saturation and therefore were limited by an upper bound. To summarize our findings and to help guide the design of the PD equivalence study, we propose a decision tree (Fig. 7) to guide the choice of PD endpoints and statistical methods based on the patient population, dose regimen, PD variability, and saturation.

If the PD similarity assessment is to be conducted in healthy subjects using a single-dose study design, it is recommended to monitor PD for a sufficiently long period, which covers the entire PD profile from onset to return to baseline. If the monitoring time is long enough, the baseline-normalized PD endpoint such as $BN\_AUEC_{(0\text{-}1y)}$ may still follow near-log-normal distribution since the distribution of the PD readout is not significantly influenced by the upper

bound of complete inhibition. Such baseline-normalized PD endpoints are suitable to be evaluated using the standard BE test.

If the PD similarity assessment is to be conducted in patients with a multiple-dose regimen, then typically AUEC endpoints during the first dosing cycle can be chosen. Since the clinical doses of many therapeutics are saturating the PD responses, baseline-normalized AUEC during the first dosing cycle may no longer follow a log-normal distribution due to an upper bound of complete inhibition. In such cases, our work has shown that the standard BE test may not be the best option due to a potentially inflated type I error. The bootstrap method, being a non-parametric method, is able to maintain a low type I error but suffered a problem of low power due to a wider 90% CI. The reason for this wider interval may be due to the small sample size to be used with replacement. The ANCOVA method, on the other hand, produces a low type I error with sufficient power and therefore can be recommended as the most suitable method for equivalence test of PD markers that has a saturable inhibition PD profile and high variability at baseline.

REFERENCES

1. Jurczak W, Moreira I, Kanakasetty GB, Munhoz E, Echeveste MA, Giri P, et al. Rituximab biosimilar and reference rituximab in patients with previously untreated advanced follicular lymphoma (ASSIST-FL): primary results from a confirmatory phase 3, double-blind, randomised, controlled study. Lancet Haematol. 2017;4:e350–61.
2. Smolen JS, Cohen SB, Tony H-P, Scheinberg M, Kivitz A, Balanescu A, et al. A randomised, double-blind trial to demonstrate bioequivalence of GP2013 and reference rituximab combined with methotrexate in patients with active rheumatoid arthritis. Ann Rheum Dis. 2017;76:1598–1602.
3. Zhu P, Ji P, Wang Y. Using clinical PK/PD studies to support no clinically meaningful differences between a proposed biosimilar and the reference product. AAPS J. 2018;20:89.
4. Putnam WS, Prabhu S, Zheng Y, Subramanyam M, Y-MC W. Pharmacokinetic, pharmacodynamic and immunogenicity comparability assessment strategies for monoclonal antibodies. Trends Biotechnol. 2010;28:509–16.
5. Evans C, Cipolla D, Chesworth T, Agurell E, Ahrens R, Conner D, et al. Equivalence considerations for orally inhaled products for local action—ISAM/IPAC-RS European workshop report. J Aerosol Med Pulm Drug Deliv. 2012;25:117–39.
6. Zhang M, Yang J, Tao L, Li L, Ma P, Fawcett JP. Acarbose bioequivalence: exploration of new Pharmacodynamic parameters. AAPS J. 2012;14:345–51.
7. Sutjandra L, Rodriguez RD, Doshi S, Ma M, Peterson MC, Jang GR, et al. Population pharmacokinetic meta-analysis of Denosumab in healthy subjects and postmenopausal women with osteopenia or osteoporosis. Clin Pharmacokinet. 2011;50:793–807.
8. Wu AHB. Biological and analytical variation of clinical biomarker testing: implications for biomarker-guided therapy. Curr Heart Fail Rep. 2013;10:434–40.
9. Trouvin A-P, Jacquot S, Grigioni S, Curis E, Dedreux I, Roucheux A, et al. Usefulness of monitoring of B cell depletion in rituximab-treated rheumatoid arthritis patients in order to predict clinical relapse: a prospective observational study. Clin Exp Immunol. 2015;180:11–8.
10. Melani L. Efficacy and safety of ezetimibe coadministered with pravastatin in patients with primary hypercholesterolemia: a prospective, randomized, double-blind trial. Eur Heart J. 2003;24:717–28.
11. Chung Chow S, Endrenyi L. Statistical issues in bioavailability/bioequivalence studies. J Bioequivalence Bioavailabil [Internet]. 2011 [cited 2018 Sep 4];01. Available from: https://www.omicsonline.org/statistical-issues-in-bioavailability-bioequivalence-studies-jbb.S1-007.php?aid=2321. Accessed 10 Oct 2018.
12. Varki R, Pequignot E, Leavitt MC, Ferber A, Kraft WK. A glycosylated recombinant human granulocyte colony stimulating factor produced in a novel protein production system (AVI-014) in healthy subjects: a first-in human, single dose, controlled study. BMC Clin Pharmacol. 2009;9(2).
13. Waller CF, Bronchud M, Mair S, Challand R. Comparison of the pharmacodynamic profiles of a biosimilar filgrastim and Amgen filgrastim: results from a randomized, phase I trial. Ann Hematol. 2010;89:971–8.
14. Liao JJ, Li Y, Jiang X. Comparability of pharmacodynamics profiles with an application to a biosimilar study. J Biometrics Biostatist [Internet]. 2017 **[cited 2018 Oct 2];08. Available from:** https://www.omicsonline.org/open-access/comparability-of-pharmacodynamics-profiles-with-an-application-to-abiosimilar-study-2155-6180-1000345.php?aid=89155Accessed 10 Oct 2018.
15. Zhu P, Sy SKB, Skerjanec A. Application of Pharmacometric analysis in the Design of Clinical Pharmacology Studies for biosimilar development. AAPS J. 2018;20:40.
16. Eastell R, Szulc P. Use of bone turnover markers in postmenopausal osteoporosis. Lancet Diab Endocrinol. 2017;5:908–23.
17. van Schaick E, Zheng J, Ruixo JJP, Gieschke R, Jacqmin P. A semi-mechanistic model of bone mineral density and bone turnover based on a circular model of bone remodeling. J Pharmacokinet Pharmacodyn. 2015;42:315–32.
18. Marathe A, Peterson MC, Mager DE. Integrated cellular bone homeostasis model for Denosumab pharmacodynamics in multiple myeloma patients. J Pharmacol Exp Ther **June**. 5(326):555–62.
19. Smolen JS, Cohen SB, Tony H-P, Scheinberg M, Kivitz A, Balanescu A, et al. A randomised, double-blind trial to demonstrate bioequivalence of GP2013 and reference rituximab combined with methotrexate in patients with active rheumatoid arthritis. Ann Rheum Dis. 2017;76:1598–602.
20. Mahmoud HK, El Nahas Y, Abdel Moaty M, Abdel Fattah R, El Emary M, El Metnawy W. Kinetics of BCR-ABL transcripts in Imatinib Mesylate treated chronic phase CML (CPCML), a predictor of response and progression free survival. Int J Biomed Sci. 2009;5:223–8.
21. Zheng J, van SE, Wu LS, Jacqmin P, Ruixo JJP. Using early biomarker data to predict long-term bone mineral density: application of semi-mechanistic bone cycle model on denosumab data. J Pharmacokinet Pharmacodyn. 2015;42:333–47.
22. FDA. Prolia (denosumab) label [Internet]. [cited 2018 Sep 4]. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2011/125320s5s6lbl.pdf. Accessed 10 Oct 2018.
23. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med. 2011;26:192–6.
24. Julious SA. Sample sizes for clinical trials with Normal data. Stat Med. 2004;23:1921–86.
25. FDA. Guidance for industry: clinical pharmacology data to support a demonstration of biosimilarity to a reference product [Internet]. 2017. Available from: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm397017.pdf. Accessed 10 Oct 2018.
26. Schrock R. Cell-based potency assays: expectations and realities. Bioprocess J. 2012;11:4–12.
27. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol. 2001;1:6.

28. Ling Z, Kun H. How to analyze change from baseline: absolute or percentage [internet]. studylib.net. **[cited 2018 Sep 12]. Available from:** http://studylib.net/doc/8421198/how-to-analyze-change-from-baseline%2D%2Dabsolute-or-percentage. Accessed 10 Oct 2018.
29. Harbeck N, Lipatov O, Frolova M, Udovitsa D, Topuzov E, Ganea-Motan DE, et al. Randomized, double-blind study comparing proposed biosimilar LA-EP2006 with reference pegfilgrastim in breast cancer. Future Oncol. 2016;12:1359–67.
30. Puri A, Niewiarowski A, Arai Y, Nomura H, Baird M, Dalrymple I, et al. Pharmacokinetics, safety, tolerability and immunogenicity of FKB327, a new biosimilar medicine of adalimumab/Humira, in healthy subjects. Br J Clin Pharmacol. 2017;83:1405–15.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.