Research Article

In Silico Evaluation of the Potential Impact of Bioanalytical Bias Difference between Two Therapeutic Protein Formulations for Pharmacokinetic Assessment in a Biocomparability Study

Theingi M. Thway,^{1,3} Chris Macaraeg,¹ Michael Eschenberg,² and Mark Ma¹

Received 17 November 2014; accepted 10 February 2015; published online 5 March 2015

Abstract. Formulation changes at later stages of biotherapeutics development require biocomparability (BC) assessment. Using simulation, this study aims to determine the potential effect of bias difference observed between the two formulations after spiking into serum in passing or failing of a critical BC study. An ELISA method with 20% total error was used to assess any bias differences between a reference (RF) and test formulations (TF) in serum. During bioanalytical comparison of these formulations, a 9% difference in bias was observed between the two formulations in sera. To determine acceptable level of bias difference between the RF and TF bioanalytically, two *in silico* simulations were performed. The *in silico* analysis showed that the likelihood of the study meeting the BC criteria was >90% when the bias difference between RF and TF in serum was 9% and the number of subjects was ≥ 20 per treatment arm. An additional simulation showed that when the bias difference was increased to 13% and the number of subjects was <40, the likelihood of meeting the BC criteria decreased to 80%. The result from *in silico* analysis allowed the bioanalytical laboratory to proceed with sample analysis using a single calibrator and quality controls made from the reference formulation. This modeling approach can be applied to other BC studies with similar situations.

KEY WORDS: bias; biocomparability; bioequivalent studies; biosimilar; biotherapeutics; PK parameter.

INTRODUCTION

During drug development, bioequivalence (BE) studies for small-molecule therapeutics and biocomparability (BC) studies for large-molecule therapeutic proteins (TPs) are often conducted to compare a new formulation (for possible commercial use) or to compare a generic or biosimilar drug to marketed drug (1). The current BE criterion per regulatory guidance mandates that the 90% confidence interval (CI) of the relative mean for $C_{\max}$ and AUC (EMA's guidance) or $C_{\max}$, AUC (0-inf) and AUC (0-t) (FDA's guidance) of the test formulation to the reference formulation should be within the range of 0.80–1.25 (2,3). For large-molecule TPs, an absence of specific guidance for BC studies has led the industry to adapt the same BE criteria wherein pharmacokinetic (PK) parameters are assessed to demonstrate therapeutic equivalence or similarity (1). Nonetheless, failure of a BE or BC study

can pose scientific challenges and financial burden to a company.

Thus, it is important for a bioanalytical scientist to understand the sources of variability that could potentially contribute to failure of a study to meet the required criteria. This understanding would allow efficient planning before initiation of sample analysis and minimize potential risks caused by bioanalytical biases. The sources of variability to be considered during planning of a BE/BC study include inter-subject variability (V^s), variability from formulation difference (V^f), and the variability of a bioanalytical method (V^m) intended to analyze serum/plasma samples collected for PK assessment from the study. V^s greater than 30% is normally considered high inter-subject variability, and its impact on designing BE/BC studies must be considered (4). While it is difficult to predict true V^s due to their physiological differences among healthy volunteers participating in a particular BC study, this may be controlled by ensuring that a sufficient number of subjects are enrolled in the study. Formulation differences (V^f) can potentially affect *in vivo* absorption, distribution, and clearance phases. The goal of such BE/BC studies is to identify V^f ; consequently, and all other sources of variability must be minimized. Additionally, method-associated variability V^m also needs to be elucidated. V^m itself has two components of variability: variability associated with random error (V^{1p}) and that associated with systematic error (V^{1n}).

¹ Department of Pharmacokinetics and Drug Metabolism, Amgen Inc., One Amgen Center Drive, Thousand Oaks, California 91320, USA.

² Department of Medical Sciences-Biostatistics, Amgen Inc., One Amgen Center Drive, Thousand Oaks, California 91320, USA.

³ To whom correspondence should be addressed. (e-mail: tthway@amgen.com)

All of the above described variables can be represented as follows:

$$\text{DataVariability}(V) = V_s + V_f + V_m = V_s + V_f + V(In + Ip) = V_s + V_f + Vip, \text{ if } VIn \leq x$$

V_s = Inter-Subject Variability and physiological component

V_f = Variability due to formulation difference – V_m = Method Variability

VIn = Inaccuracy or bias,

VIp = Imprecision,

If the bioanalytical method has been validated, the total error (TE) of the method, which includes random errors (V_{Ip}) as assessed by imprecision or inter-assay %CV and systematic errors (V_{In}) as assessed by inaccuracy or %bias, has been established during the pre-study method validation (2,3). While the systematic bias of a particular TP in serum can be due to various formulation-related factors, the bias difference (ie differences in V_{In}) between reference formulation (RF) and test formulation (TF) may further affect the accurate measurement of the analyte in the case of BC studies. The term “bias difference” in this report referred to the absolute difference of V_{In} observed between two formulations in 100% serum during bioanalytical comparability testing rather than the bias of two formulations during analytical testing. In a typical non-BC PK study, only one formulation is used to dose subjects and to spike in serum pool for preparation of a standard calibrator. As a result, the systematic error of the bioanalytical method used for sample analysis can be minimized using a properly evaluated bulk-prepared standard calibrator once long term-storage stability is established. In a BC study, the additional test formulation may contribute to inaccuracy (V_{In}) that can be addressed if the bias difference between TF and RF can be determined prior to the bioanalysis. If no bias difference between the RF and TF is observed, it can be assumed that V_{In} is equal to “0” and that one bulk-prepared standard calibrator can be used to analyze the samples collected from subjects treated with either formulation to further reduce V_{Ip} . However, other sources of V_{Ip} that are difficult to control bioanalytically include day-to-day, plate-to-plate, and analyst-to-analyst variability.

To minimize all sources of variability and to successfully support the BC study, a bioanalytical comparability test must be performed between the two formulations.

Figure 1 represents the process flow for determining the bioanalytical comparability between two formulations: in the first scenario, when a validated bioanalytical method is used to evaluate bioanalytical comparability and in the second, where the analytical method has not been validated. In scenario one, the bias difference between RF and TF can be determined bioanalytically by performing three to six analytical runs by two to three analysts. Each run shall include a set of standard calibrator (STD) and two sets of five levels of quality controls (QCs) including the lower limit of quantification (LLOQ) and upper limit of quantification (ULOQ),

each prepared using both formulations. The TE, bias, and %CV of TF must meet the method acceptance criteria. Additionally, the TE, bias, and %CV between RF and TF should be separately and collectively calculated and assessed. Specifically, TE, bias, and %CV of each formulation QC should be determined against STDs of RF and TF. If available, historical data for variability-associated bias can be used to tease out the bias difference between RF and TF. In the second scenario, where a validated method is not available, both RF and TF must be validated through accuracy and precision runs during pre-study method validation, and the bias difference must be identified in a similar manner to that used in the first scenario. The method acceptance criteria can be defined using the cumulative data from both formulations. If the bias difference between two formulations (ie, bias of TF QC against RF STD and vice versa) in 100% matrix is found to be <5% during the bioanalytical comparison, then it is reasonable to analyze each subject profile from each cohort (two subject profiles, treated with RF or TF) in a single plate using a standard calibrator prepared from RF to minimize plate-to-plate variation. Given the fact that LBA typically has larger inherent variability, the question being raised here is the acceptable bias difference between two formulations for the study to meet BC criterion in the presence of inter-subject variability. Specifically, in the presented study, if a method TE is 20% and subject variability is around 30%, it is important to know the tolerable bias difference between two formulations to meet the current criterion for the BC study. The TF met the same specifications of RF indicating that the two formulations were “expected” to behave similarly. However, during the bioanalytical comparability test, we had observed the bias difference ranging from 4 to 9%. We hypothesize that if the bias difference between two formulations exceeds 10%, the chances of failing the BC study may increase. The goal of this simulation is to determine the acceptable level of bias difference between two formulations in 100% serum. The analysis aims to investigate the likelihood that V_m may mask true formulation differences and impact passing or failing of the critical BC study if there is a known bioanalytical bias difference between two formulations.

This finding provides an appropriate course of action for a bioanalytical strategy, allowing an RF to be used as the standard calibrator in all runs during sample analysis, and the bias difference between two formulations ranges from 9 to

Fig. 1. Prerequisite process in establishing bioanalytical comparability for biocomparability studies. *RF* Reference formulation; *TE* Total error; *TF* Test formulation

13%. The simulation creates a framework from which to develop a realistically designed bioanalytical strategy for the actual BC study.

MATERIALS AND METHODS

Bioanalytical Method for Measurement of Mab-1 Concentration

STD and QC were prepared by spiking a monoclonal therapeutic antibody (Mab-1) (Lot #785A078089 or 0010081396) in 100% normal human serum and stored at $-70^{\circ}\text{C} \pm 10^{\circ}\text{C}$. Mab-1 is a humanized IgG₂ antibody and formulated at the nominal concentration of 70 mg/mL. Microplates (96 wells) were coated with 2 $\mu\text{g/mL}$ of a murine antihuman IgG monoclonal antibody against Mab-1 (Clone no. 1, Amgen, Inc. CA) in sodium bicarbonate buffer. Plates were blocked with 300 μL of I-block buffer. The STD, three levels of QC, and blank were diluted 1:50 in assay buffer (1X PBS with 0.5 M NaCl and 0.5% Tween 20) before loading, and 100 μL of each was added into a 96-well microtiter plate. After a wash step, a 100 μL of HRP-conjugated mouse antihuman IgG monoclonal antibody against Mab-1 (Clone 2, Amgen Inc., CA) at the approximate concentration of 5 ng/mL was added to the wells. TMB peroxide substrate solution (Kirkegaard and Perry Laboratories Inc., MD) was added to the wells after the final wash step. Color development was stopped with acidification, and the optical density (OD) was measured at 450 nm with reference to 650 nm using Molecular Devices Spectramax 340PC microplate reader (Molecular Devices, Sunnyvale, CA) equipped with SOFTmax Pro Version 5.0.1 software. The assay dynamic range of the analytical method was 50 to 2495 ng/mL.

Formulations

The RF with a 70-mg/mL nominal concentration was used in Phase I to Phase III studies, while the TF with a concentration >200 mg/mL intended for future late stage studies was tested.

Method Validation

Accuracy and precision experiments were conducted by performing a total of six runs on three different days by three different analysts. The STD concentrations were 25, 50, 100, 150, 200, 312, 624, 1248, 1997, 2492, and 4993 ng/mL. Five level QCs at the LLOQ, LQC (low level QC), MQC (midlevel QC), HQC (high level QC), and ULOQ were prepared by spiking the Mab-1 into

prescreened and 100% pooled human serum. These five levels QCs were called validation samples (VSs) during the accuracy and precision runs. All analysts prepared their own set of STDs, and a single analyst prepared VSs in bulk for all six runs. The accuracy (% bias), precision (% coefficient of variance [CV]), and TE (sum of % bias and % CV) performance characteristics were calculated using a validated ligand-binding assay's EXCEL Software (Amgen Inc., Thousand Oaks, CA) based on the white paper described by DeSilva *et al.* (2).

Comparison of RF to TF in 100% Serum for Assessment of Bias Difference

The method specified above was validated and was used to support multiple clinical studies. The same method was used to qualify the TF lot in support of a specific BC study. For the TF qualification and bias evaluation, two independent preparations of STDs and five levels of QC sets were prepared from each formulation (both RF and TF) in 100% human serum using a Tecan EVO[®] Freedom (Tecan, Männedorf, Switzerland) workstation. The five levels of QCs included were LLOQ, LQC, MQC, HQC, and ULOQ. Six analytical runs were performed by three analysts, two runs per analyst on three different days. Each analytical run included a set of STDs from each formulation and two sets of five levels of QC from each formulation.

To qualify TF using the bioanalytical method, the QCs spiked with TF in 100% serum were evaluated against the TF-prepared STDs. To assess the bias, the QCs prepared with either formulation were evaluated against both the RF and TF STD. The intra- and inter-assay performance characteristics and TE (sum of % bias and %CV) for each formulation were determined using a validated ligand-binding assay's EXCEL Software (Amgen Inc., Thousand Oaks, CA) based on the white paper described by DeSilva *et al.* (2).

Study Design

A randomized, double-blind, placebo-controlled, ascending-single-dose study in healthy men and postmenopausal women (clinical study A) was a first-in-human single dose study. One of the cohorts in this study included a dosing regimen of 3-mg/kg subcutaneous (SC) administration, which is equivalent to a 210-mg fixed dose for subjects with a typical body weight of 70 kg. The concentration–time data

from this study were used as a reference in the simulation experiment since the same time points were used to collect samples in the BC study.

Statistical Method for Simulation and Software

Simulations were performed by first generating subject-level concentration data across 16 sample collection time points (6, 12, 24, 48, 72, 96, 120, 168, 240, 336, 504, 672, 840, 1008, 1176, and 1344 h post-dose) based on the observed mean and variability (total variability—method and subject) at each time point of a compound, assessed internally. Within each time point, the concentrations were assumed to be normally distributed across subjects. The applied variability at each time point ranged from 15 to 81% (%CV) with the later time points having the greater variability, as this is what was reported in clinical study A. BC study simulations were based on a set number of subjects ($N=20, 40, 80$, and 120) for the particular simulation, with half of the subjects assigned to a unique bias (Bias 1) and the other half of the subjects assigned to a different bias (Bias 2).

The assigned bias was applied to each simulated concentration where the bias was considered to be normally distributed (not equal for all subjects, and all time points for a subject assigned a given bias). The simulated concentration–time data are then used for the calculation of the AUC values for each subject. For each simulated subject, the area under the curve ($AUC_{(0-1344\text{ h})}$) was computed using the linear trapezoidal rule based on the simulated (biased) concentration–time data with the pre-dose time-point (hour 0) assigned the value of 0. No extrapolation of AUC was performed ($AUC_{(0-1344\text{ h})}$ was calculated). The AUC values were then log-transformed and compared using two one-sided tests to evaluate equivalence between the two bias levels (Bias1 vs. Bias2). Two one-sided tests are employed and evaluated using PROC MIXED within SAS version 9.2 (SAS, Cary, NC). The results were then back transformed to give the estimated ratio of the geometric mean and the 90% CI of the ratio for AUC. The AUC for the two bias levels was declared equivalent (passed) if the 90% CI of the ratio was completely contained within the equivalence range 0.80–1.25. This process was followed, and the analysis was performed 1000 times for each set of bias levels and for each sample size. Each iteration contained a unique set of randomly generated data. The passing rate was then evaluated for each of the sample size, difference in bias, and specific level of bias combinations.

RESULTS

Accuracy and Precision of the Method during Bioanalytical Method Validation

Following the successful runs of the six accuracy and precision experiments, a cumulative assessment for accuracy and precision was performed. The accuracy and precision of STDs ranged from -2 to 8% and 2 to 6% , respectively. The inter-assay accuracy for five level QCs (LLOQ, LQC, MQC, HQC, and ULOQ) ranged from -9 to 3% , and inter-assay precision ranged from 7 to 13% . The TE of the method was determined as $\leq 17\%$ (data not shown). Based on these accuracy and precision results, method acceptance criterion for QC was set at 20% TE for the study support.

Comparison of RF to TF in 100% Serum for Assessment of Bias Difference

Before supporting sample analysis for a BC study, the qualification of TF to RF in 100% serum was conducted following the processes described in Fig. 1. The differences in bias between the two formulations were determined from the six formulation qualification runs (step 2 in Fig. 1). When the STD calibrator prepared with RF was used to calculate the recoveries of QCs from each formulation, the inter-assay accuracy for QCs ranged from -1.9 to 0.5% and -5.1 to -3.0% for RF and TF, respectively (Fig. 2a). The inter-assay precision for QCs ranged from 4 to 6% and 4 to 11% for RF and TF, respectively (Fig. 2a). The TE for QCs was $<8\%$ for RF and $<16\%$ for TF. If the STD calibrator prepared with TF was used to calculate the recoveries of each formulation QC, the inter-assay bias ranged from -0.3 to 5.5% and -3.2 to 3.1% for TF and RF, respectively (Fig. 2b). The inter-assay precision for QCs ranged from 5 to 7% and 2 to 7% for TF and RF, respectively. The TE for QCs was $<11\%$ for both RF and TF (Fig. 2b). Thus, the maximal bias difference among QCs between RF and TF was 3.9 and 8.6% against RF and TF, respectively (Fig. 2c).

Assessment of Bias Effect in Simulated AUC

To further understand the effect of bioanalytical bias difference between the two formulations, the contribution of inter-subject variability was considered in addition to method variability. It is considered that there is little or no inter-subject variability if the inter-subject %CV is $<30\%$ (4). Based on this criterion, there is no inter-subject variability if the Mab-1 concentrations were >3000 ng/mL (quadrant III) obtained from clinical study A (Fig. 3a). Higher inter-subject variability was observed for Mab-1 concentration <2000 ng/mL (quadrant I). Therefore, the concentration-dependent inter-subject variability was taken into consideration during the simulation study.

Additionally, the acceptance criterion for the estimated ratio of the geometric mean and the 90% confidence interval (CI) for the exposure area AUC ratio was set at 0.80 to 1.25 so as to be considered equivalent. Since the maximum observed bias difference between QCs for RF and TF was 8.6% , the passing and failing rates in achieving BC criteria were predicted based on various bias combinations with 9% of maximal difference between RF and TF in 100% serum (Fig. 4a) with a varying number of subjects. Additionally, the passing and failing rates were predicted on various bias combinations with larger bias difference of 13% (Fig. 4b). Both analyses were done for different numbers of subjects ($N=20, 40, 80$, and 120), where half was assigned Bias 1, and the other half was assigned Bias 2. The combination bias pairs of RF and TF ranged from -10 to 9 for 9% in bioanalytical bias difference and from -12 to 13 for 13% in bioanalytical bias difference.

When the absolute bias difference is 9% (Fig. 4a), the likelihood of achieving BC criteria was $>85\%$ for all four bias combinations; a ($-10, 1$), b ($-5, 4$), c ($-4, 5$), and d ($0, 9$) even when the total subject number is 20. When the total subject numbers ≥ 40 (i.e., $N \geq 20$ per treatment arm), the likelihood of achieving BC criteria was still $>95\%$ even when the absolute bias combination is $>9\%$ i.e., a ($-10, 1$). In contrast,

Fig. 2. Qualification of test formulation (TF) using reference formulation (RF) in six qualification runs. Standard calibrator and five level QCs were prepared using RF and TF. Percent bias and inter-assay %CV for both formulations were measured using standard curves from RF (a) or TF (b). *LLOQ* lower limit of quantification, *LQC* low QC, *MQC* mid QC, *HQC* high QC, *ULOQ* upper limit of quantification. The maximal bias difference between RF and TF was calculated by combining the %Bias of RF and TF in all QC levels when the QC was measured against either RF or TF (c)

when the absolute bias difference is 13% (Fig. 4b) with all four bias combinations: a (−12, 1), b (−7, 6), c (−6, 7), and d (0, 13), the likelihood of achieving BC criteria was <80% when the subject number was 20. When the total subject

numbers ≥ 20 , the likelihood of achieving BC criteria was still >95% even when the absolute bias combination is 13% *i.e.*, a (−12, 1) and d (0, 13).

DISCUSSION

BC studies are pivotal for drug development and are a critical component for ensuring that the patients receive the same quality and exposure of therapeutic, which in turn should guarantee efficacy and safety. In this study, we performed a formulation qualification between RF and TF in 100% human serum prior to supporting a BC study where PK was the primary endpoint. A nearly 9% difference in bias was observed during the formulation qualification that prompted us to evaluate the potential impact that this formulation bias may have on BC study meeting the acceptance criteria. A simulation study was conducted to determine the proper approach to supporting the bioanalytical portion of the BC study. This information was critical to how the study should be conducted so as to minimize any bioanalytical concerns that could influence the PK results. In most cases, the RF and TFs are bioanalytically comparable which allows the study to be conducted using a standard calibrator prepared with RF for sample analysis. In order to minimize analytical method variability, various

Fig. 3. Inter-subject variability ($N=6$) of Mab-1 administered at 70-mg/kg dose into healthy volunteers from Study A plotted against the mean serum concentrations per collection time point. Variability vs concentration plot was divided into four quadrants (I, II, III, and IV)

◀ **Fig. 4.** Simulation evaluation of bias combinations with respect to passing/failing of biocomparability criteria for a given study where the bias difference between reference and test formulations is 9% (a) and 13% (b), the number of subjects in each simulated study is indicated. Bias combinations in x-axis indicate the Bias 1 and Bias 2 percentages simulated. The y-axis indicates the probability of the biocomparability study meeting and passing (blue) or failing (red) bioequivalent criteria for the given sample size and bias combination

control measures such as analysis of an entire profile per subject in the same plate, shortening the sample analysis period, training all the analysts to perform consistently and similarly are implemented. To further prevent plate-to-plate and run-to-run variability, two subject profiles, one from each treatment arm, were evaluated on the sample assay plate. Even after these measures, bioanalytical method and subject-to-subject variability remained the most plausible sources of variability. However, if the two formulations in 100% serum were found to have a larger bias difference during bioanalytical comparison testing, this can further increase the risk of failure of BC study due to increased data variability. In the current case study, nearly, a 9% difference in bias was observed during the formulation qualification. The 9% variation in V_{In} would leave only a little room for V_{Ip} and V_s when the LBA method with 20% TE was used, thus true V_f may be masked. Simulation allowed us to understand the effect of bias difference and assist in the strategic planning during the bioanalytical phase of the study. Additionally, when possible, study design includes sufficient subject numbers so that results can be statistically interpretable. In this case, the planned BC study included sufficient subjects ($N=80$) in each treatment arm. Based on the simulation result, it is reasonable to believe that a single standard calibrator prepared from RF can be used to analyze samples from subjects treated with each formulation without compromising the resultant data. Although the data from the simulation study cannot be generalized and directly applied to the other BC studies conducted with different therapeutics and/or with similar issues where V_s , V_{In} and V_{Ip} may differ, this study provided the unique approach that bioanalytical simulation could provide and allow scientists to proactively evaluate and plan. Additionally, *in silico* analysis showed that lesser subject numbers ($N=40$) could still meet the

BC criteria even with 13% bias difference between RF and TF in serum and 20% TE method.

CONCLUSIONS

BE and BC studies are necessary and critical for the drug development companies to evaluate new formulations and/or improve the manufacturing processes to ensure drug safety and efficacy. Success of a BE or BC study requires that each component in the study must be thoroughly examined during the planning phase. The bioanalytical simulation conducted illuminated the potential impacts on a study if the bias difference between RF and TF is higher than expected. The simulation result allowed a suitable evaluation of bioanalytical strategy in supporting the sample analysis for the comparability study. Finally, this modeling approach can be applied to other BC studies with similar situations.

ACKNOWLEDGMENTS

Financial support for this research was provided by Amgen Inc. All authors own Amgen stocks. We would like to thank Dr. Michael Hall and Dr. Vibha Jawa (both from Amgen Inc) for their critical review and Dr Jeffery Talbot for providing editorial support.

REFERENCES

1. Wang D *et al.* Clinical trials a practical guide to design, analysis and reportings. Chapter 13 bioequivalence trials.
2. DeSilva B *et al.* Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules. *Pharm Res.* 2003;20:1885–900.
3. Thway TM *et al.* Applications of a planar electrochemiluminescence platform to support regulated studies of macromolecules: benefits and limitations in assay range. *J Pharm Biomed Anal.* 2010;51:626–32.
4. Van Peer A. Variability and impact on design of bioequivalence studies. *Basic Clin Pharmacol Toxicol.* 2010;106:146–53.