

## ON STATISTICAL EVALUATION OF THE LINEARITY IN ASSAY VALIDATION

Eric Hsieh<sup>1</sup> and Jen-pei Liu<sup>1,2</sup>

<sup>1</sup>Division of Biometry, Institute of Agronomy, National Taiwan University, Taipei, Taiwan

<sup>2</sup>Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan, Taiwan

*Linearity is one of the most important characteristics for evaluation of the accuracy in assay validation. The current statistical method for evaluation of the linearity recommended by the Clinical Laboratory Standard Institute (CLSI) guideline EP6-A is reviewed. The method directly compares the point estimates with the pre-specified allowable limit and completely ignores the sampling error of the point estimates. An alternative method for evaluation of linearity, proposed by Kroll et al. (2000), considers the statistical test procedure based on the average deviation from linearity (ADL). However this procedure is based on an inappropriate formulation of hypotheses for the evaluation of linearity. Consequently, the type I error rates of both current methods may be inflated for inference of linearity. To claim the linearity of analytical methods, we propose that the hypothesis of proving the linearity should be formulated as the alternative hypothesis. Furthermore, any procedures for assessment of linearity should be based on the sampling distributions of the proposed test statistics. Therefore, we propose a two one-sided test (TOST) procedure and a corrected Kroll's procedure. The simulation studies were conducted to empirically compare the size and power between current and proposed methods. The simulation results show that the proposed methods not only adequately control size but also provide sufficient power. A numeric example illustrates the proposed methods.*

**Key Words:** Allowable limit; Linearity; Quantitative analytical laboratory methods.

### INTRODUCTION

Linearity relationship is one of the most important characteristics of accuracy for validation of quantitative analytical laboratory procedures. According to the ICH Q2A guideline (ICH, 1995), the linearity of an analytical method is defined as its ability (within a given range) to obtain the test results, which are directly proportional to the concentration (amount) of the analyte in the test sample. Evaluation of linearity whether a mathematically verified straight-line relationship exists between the observed values and the true concentrations or activities of the analyte (ICH, 1996). Linearity is important for analytical and clinical laboratory

Received July 21, 2007; Accepted September 3, 2007

Address correspondence to Jen-pei Liu, Division of Biometry, Department of Agronomy, National Taiwan University, 1, Section 4, Roosevelt Road, Taipei, Taiwan; E-mail: jpliu@ntu.edu.tw

methods because it represents the simplest mathematical relationship and allows for simple and easy interpolations of results for clinical practitioners. (Schofield, 2003a,b; Zhong et al., 2003) Although the four-parameter logistic equation is commonly used for radioimmunoassay (RIA), which employs the principle of antibody-antigen as a means to quantify an analyte, we focus analytical methods described in the approved Clinical Laboratory Standard Institute (CLSI) guideline EP6-A. (CLSI, 2003)

According to the approved CLSI guideline EP6-A, an experiment for evaluation of linearity should include at least five solutions of different concentration levels across the anticipated range. Two to four replicates should be run at each level. The guideline suggests that linearity is claimed if the difference between the best-fit nonlinear polynomial curve and simple linear regression equation at each concentration is less than the pre-defined allowable bias. An alternative method for evaluation of linearity, proposed by Kroll et al. (2000), considers the statistical test procedure based on the average deviation from linearity (ADL), defined as the square-root of the average squared distances between the fitted point on the best fit polynomial curve and the simple regression equation at each solution level, standardized by mean concentration. The sampling distribution of the observed ADL is a function of a noncentral Chi-square distribution with the noncentrality parameter defined as some pre-specified allowable limit. Linearity is concluded at the  $\alpha$ -nominal level if the observed value of the ADL is smaller than the  $(1 - \alpha)\%$  critical value obtained from the sampling distribution of the observed ADL.

However, the procedure for assessment of linearity, based on ADL proposed by Kroll et al. (2000), is derived from formulating a hypothesis for proving linearity as the null hypothesis. In addition, the critical values used in determination of linearity suggested in Kroll et al. (2000) contains unknown parameters that are estimated by sample statistics. On the other hand, the method for evaluation of linearity recommended by EP6-A directly compares the point estimates with the pre-specified allowable limit and completely ignores the sampling error of the point estimates. As a result, the type I error rate may be inflated and the probability of incorrect claims of linearity is not adequately controlled at the nominal level. We argue that the hypothesis of proving the linearity should be formulated as the alternative hypothesis. Furthermore, any procedures for assessment of linearity should be based on the sampling distribution of the proposed test statistics. Therefore, we propose a two one-sided test (TOST) procedure and a corrected Kroll's procedure for assessing linearity in validation of quantitative analytical laboratory procedures. We also review current methods for the evaluation of linearity, including the procedures suggested in the approved CLSI guideline EP6-A (2003) and by Kroll et al. (2000). The shortcomings of these methods are highlighted. In addition, a two one-sided test (TOST) procedure and a corrected Kroll's procedure for assessing the linearity of analytical methods are also proposed; the results of a comprehensive simulation study for comparison of empirical size and power of proposed methods with the current methods are provided. Numeric examples are then employed to illustrate and compare the current and proposed methods. Discussion and final remarks are provided.

**CURRENT METHODS FOR EVALUATION OF LINEARITY**

The approved CLSI guideline EP6-A (2003) recommends that the experiment for linearity assessment should be conducted on at least five solutions of different concentrations run at least in duplicates. Let  $Y_{ij}$  be the test result of replicate  $j$  at concentration  $X_i$ , where  $j = 1, \dots, R$ ;  $i = 1, \dots, L$ . The approved CLSI guideline EP6-A considers the following linear, quadratic, and cubic models fitting the data obtained from the experiment:

$$\begin{aligned} \text{Linear (first order)} \quad \mu_{Li} &= \alpha' + \beta'_1 X_i \\ \text{Quadratic (second-order polynomial)} \quad \mu_{Qi} &= \alpha'' + \beta''_1 X_i + \beta''_2 X_i^2 \quad \text{or} \\ \text{Cubic (third-order polynomial)} \quad \mu_{Ci} &= \alpha''' + \beta'''_1 X_i + \beta'''_2 X_i^2 + \beta'''_3 X_i^3 \quad (1) \end{aligned}$$

where  $\mu_{Li}$ ,  $\mu_{Qi}$ , and  $\mu_{Ci}$  are the predicted means of the corresponding models and  $\alpha'$ ,  $\alpha''$ ,  $\alpha'''$ ;  $\beta'_1, \beta''_1, \beta'''_1$ ;  $\beta''_2, \beta'''_2$ , and  $\beta'''_3$  are the intercepts, regression coefficients for the corresponding models in Eq. (1). In what follows, all assumptions for fitting the best-fit model specified in the approved CLSI guideline EP6-A are satisfied and we use the definition of the best-fitted model recommended by EP6-A. The best-fit model is the model where the result of lack-of-fit test is not statistically significant and repeatability meets the manufacturer’s claim. In addition, for the purpose of illustration, we also assume that random error is approximately constant rather than proportional in the range of concentrations considered in the experiment. If the best-fitted model is the linear model over the some range of concentrations employed in the experiment, then the assay method can be concluded to be linear over the some range of concentrations. However, if the best-fitted model is not linear, the linearity of the analytical procedure can still be claimed if the magnitude of deviations from linearity at each concentration is within some pre-specified allowable limit of  $\delta_0$ .

Let  $\mu_{Pi}$  be the predicted mean of the best-fitted model, where  $\mu_{Pi}$  can be either  $\mu_{Qi}$  or  $\mu_{Ci}$ . The difference in predicted means between the best-fit nonlinear and linear model  $\mu_{Pi} - \mu_{Li}$  represents a measure of the degree of deviation from linearity at each concentration level. The hypothesis for evaluation of linearity can be formulated as

$$H_{0i} : |\mu_{Pi} - \mu_{Li}| \geq \delta_0 \quad \text{vs.} \quad H_{ai} : |\mu_{Pi} - \mu_{Li}| < \delta_0, \quad \text{for all } i = 1, \dots, L. \quad (2)$$

Because hypothesis (2) requires all differences in the predicted means between the best-fitting and linear models to be within the pre-specified allowable limit, it is a disaggregate criterion (Chow and Liu, 2000). Let  $\widehat{Y}_{Pi}$  and  $\widehat{Y}_{Li}$  be the least-squared (LS) estimators of the predicted mean of the best-fit and linear models, respectively, where

$$\begin{aligned} \widehat{Y}_{Li} &= a' + b'_1 X_i, \quad \text{and} \\ \widehat{Y}_{Pi} &= \begin{cases} a'' + b''_1 X_i + b''_2 X_i^2, & \text{if the best-fitted model is quadratic,} \\ a''' + b'''_1 X_i + b'''_2 X_i^2 + b'''_3 X_i^3, & \text{if the best-fitted model is cubic;} \end{cases} \end{aligned}$$

and  $a', a'', a'''$ ;  $b'_1, b''_1, b'''_1$ ;  $b''_2, b'''_2$ , and  $b'''_3$  are the LS estimators of the intercepts, regression coefficients for the corresponding models in (1).

According to the approved CLSI EP6-A, the linearity of the proposed analytical method can be concluded if

$$|\widehat{Y}_{Pi} - \widehat{Y}_{Li}| < \delta_0, \quad \text{for } i = 1, \dots, L. \quad (3)$$

This method is referred to as the estimation method because it only considers estimators for the evaluation of linearity. The estimation method completely ignores the variability and distribution associated with the estimators. Therefore, it also may inflate the type I error rate in assessment of linearity.

Kroll et al. (2000) considered the average deviation from linearity (ADL) for assessment of linearity. The ADL is defined as

$$\theta = \text{ADL} = \frac{\sqrt{\sum_{i=1}^L (\mu_{Pi} - \mu_{Li})^2 / L}}{\mu}, \quad (4)$$

where  $\mu$  is the population mean concentration for all solutions of the assay.

Since ADL is a function of the standardized sum of squares of the differences in the predicted means between the best-fitted and linear models, it is an aggregate criterion. Therefore, the hypothesis for evaluation of linearity based on ADL proposed by Kroll et al. (2000) is given as

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_a : \theta > \theta_0, \quad (5)$$

where  $\theta_0$  is the maximum allowable average deviation from linearity.

The value  $\theta_0$  is usually presented as a percent. Kroll et al. (2000) refer it to as percent bound and suggest 5% as a reasonable cutoff for most relevant clinical applications.

An estimator of ADL suggested by Kroll et al. (2000) is given as

$$\hat{\theta} = \frac{\sqrt{\sum_{i=1}^L (\widehat{Y}_{Pi} - \widehat{Y}_{Li})^2 / L}}{\bar{X}}, \quad (6)$$

where  $\bar{X}$  is the observed mean concentration for all solutions of the assay.

At  $\theta = \theta_0$ ,  $\sum_{i=1}^L (\widehat{Y}_{Pi} - \widehat{Y}_{Li})^2$  follows a non-central Chi-square distribution with degrees of freedom  $d - 1$ , and noncentrality parameter  $\text{LR}\theta_0^2 / (\sigma/\mu)^2$ , where  $d$  is the degrees of freedom for regression of the best-fitted model and  $\sigma^2$  is the variance of residuals under the best-fitted model.

The decision rule for concluding that the assay is linear at the 5% significance level if

$$\hat{\theta} < \frac{\sigma}{\mu} \sqrt{\frac{q_{0.95}}{\text{LR}}}, \quad (7)$$

where  $q_{0.95}$  is the 95th percentile of a noncentral Chi-square distribution with degrees of freedom  $d - 1$  and noncentrality parameter  $\text{LR}\theta_0^2 / (\sigma/\mu)^2$ .

We refer this method to as the uncorrected Kroll method.

The method proposed by Kroll et al. (2000) has two shortcomings. The hypothesis for proving linearity is formulated as the null hypothesis. When

Eq. (7) is satisfied, the only conclusion is that the null hypothesis is not rejected and this does not imply that the linearity of the assay is proved. On the other hand, the critical value in (7) contains the unknown parameters  $\mu$  and  $\sigma$ , which need to be estimated from the data. Kroll et al. (2000) suggested estimating  $\mu$  by  $\bar{X}$ , the observed mean concentration for all solutions of the assay, and  $\sigma$  by the square-root of residual mean square obtained the best-fitted model. Consequently, the variability associated with the residual mean square is not considered in evaluation of linearity by Eq. (7). Because of these two shortcomings, the method based on ADL proposed by Kroll et al. may not adequately control the type I error rate at the nominal level for evaluation of linearity.

**PROPOSED PROCEDURES**

The interval hypothesis in (2) can also be decomposed into two sets of one-sided hypotheses, as

$$\begin{aligned}
 H_{0iL} : \mu_{Pi} - \mu_{Li} \leq -\delta_0 \quad \text{vs.} \quad H_{aiL} : \mu_{Pi} - \mu_{Li} > -\delta_0, \quad \text{for all } i = 1, \dots, L, \quad \text{and} \\
 H_{0iU} : \mu_{Pi} - \mu_{Li} > \delta_0 \quad \text{vs.} \quad H_{aiU} : \mu_{Pi} - \mu_{Li} < \delta_0, \quad \text{for all } i = 1, \dots, L.
 \end{aligned}
 \tag{8}$$

An unbiased estimator of  $\mu_{Pi} - \mu_{Li}$  is the LS estimator  $\widehat{Y}_{Pi} - \widehat{Y}_{Li}$ ,  $i = 1, \dots, L$ . Define

$$\begin{aligned}
 \mathbf{X}_L &= (\mathbf{1}, \mathbf{X}), \\
 \mathbf{X}_p &= \begin{cases} (\mathbf{1}, \mathbf{X}, \mathbf{X}_2), & \text{if the best-fitted model is quadratic, and} \\ (\mathbf{1}, \mathbf{X}, \mathbf{X}_2, \mathbf{X}_3), & \text{if the best-fitted model is cubic,} \end{cases}
 \end{aligned}$$

where  $\mathbf{1}$  is  $LR \times 1$  vector of 1s,  $\mathbf{X} = (X_i)$ ,  $\mathbf{X}_2 = (X_i^2)$ , and  $\mathbf{X}_3 = (X_i^3)$ .

An unbiased estimator of the variance for  $\widehat{Y}_{Pi} - \widehat{Y}_{Li}$  is given as

$$\widehat{\sigma}_{di}^2 = w_{ii} \widehat{\sigma}_e^2,$$

where  $w_{ii}$  is the  $i$ th diagonal element of matrix  $\mathbf{W}$ ,  $i = 1, L$ ;  $\widehat{\sigma}_e^2$  is the residual mean square obtained from the best-fitted model with degrees of freedom  $LR-d-1$ ,  $\mathbf{W} = \mathbf{W}_p - \mathbf{W}_L$ ,  $\mathbf{W}_p$  and  $\mathbf{W}_L$  are the projection matrices corresponding to the column spaces spanned by the design matrices of the best-fitted and linear models, respectively, i.e.,  $\mathbf{W}_p = \mathbf{X}_p(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p$  and  $\mathbf{W}_L = \mathbf{X}_L(\mathbf{X}'_L\mathbf{X}_L)^{-1}\mathbf{X}'_L$ .

It follows that the  $(1 - 2\alpha)100\%$  confidence interval for  $\mu_{Pi} - \mu_{Li}$  is given as

$$(\widehat{Y}_{Pi} - \widehat{Y}_{Li}) \pm t_{\alpha, LR-d-1} \widehat{\sigma}_{di}, \quad i = 1, \dots, L,
 \tag{9}$$

where  $t_{\alpha, LR-d-1}$  is the upper  $\alpha$  percentile of a central distribution with a degree of freedom of  $LR - d - 1$ .

The linearity of an analytical method can be concluded at the  $\alpha$  significance level if the  $(1 - 2\alpha)100\%$  confidence interval for  $\mu_{Pi} - \mu_{Li}$  is completely contained within the pre-specified allowable limit of  $\pm\delta_0$  at each concentration level,  $i = 1, \dots, L$ . This method is referred to as the two one-sided tests (TOST) procedure.

The main drawback of the method for evaluation of linearity proposed by Kroll et al. (2000) is the incorrect formulation of the hypotheses. We suggest the hypothesis for assessment of linearity based on ADL should be formulated as follows:

$$H_0 : \theta \geq \theta_0 \quad \text{vs.} \quad H_a : \theta < \theta_0. \quad (10)$$

Consequently, the linearity of an analytical procedure is concluded at the 5% significance level if

$$\hat{\theta} < \frac{\sigma}{\mu} \sqrt{\frac{q_{0.05}}{LR}}, \quad (11)$$

where  $q_{0.05}$  is the 5th percentile of a noncentral Chi-square distribution with degrees of freedom  $d - 1$  and noncentrality parameter  $LR\theta_0^2/(\sigma/\bar{X})^2$ . This method is referred to as the corrected Kroll method.

## SIMULATION STUDY

We conducted a simulation study to compare the empirical sizes and powers of the uncorrected Kroll method, the corrected Kroll method, the estimation, and two one-sided tests procedures. Following the specification of the experimental designs for evaluation of linearity, the number of solutions (or dilutions) of different concentrations is set to be 5 or 7 and the number of replications at each concentration is 2, 3, or 4. Throughout the simulation, the allowable margin of linearity based on ADL,  $\theta_0$ , is specified as 0.05 while the margin for the estimation and TOST procedures,  $\delta_0$ , is specified as 0.2. There are two types of comparison of size. The first type is to compare the size between the uncorrected Kroll with corrected Kroll methods, for which the data were generated at the value of 0.05 for ADL as recommended by Kroll et al. (2000). The second type is to compare the size between the estimation method suggested in the approved CLSI guideline and the TOST procedure, for which the data were generated with the true difference,  $\mu_{pi} - \mu_{Ci}$  at some solutions being either 0.2 or  $-0.2$ . In addition, the standard deviation of normal random error was specified as 0.1 and 0.2. Table 1 provides the specifications of the values of parameters in the simulation for evaluation of size. For each of 12 combinations, five thousand (5000) random samples are generated. For the 5% nominal significance level, a simulation study with 5000 random samples implies that 95 percent of the empirical sizes evaluated at the equivalence limits will be within 0.04395 and 0.05604 if the proposed methods can adequately control the size at the nominal level of 0.05. A testing procedure is considered too liberal (conservative) if its estimated size is greater (smaller) than 0.05604 (0.04395). In addition, the specifications of parameters for investigation of power are given Table 2.

Table 3 presents the results of the empirical sizes. For the comparison between the uncorrected Kroll and the correct Kroll methods, all empirical sizes of the uncorrected Kroll method are above 0.92. On the other hand, the empirical size of the corrected Kroll method ranges from 0.0516 to 0.078. Only 8.33% (1/12) of the empirical sizes of the corrected Kroll method are within 0.04395 and 0.05604.

**Table 1** Specifications of parameters for size

Type of comparisons	No. of solution levels	True ADL	Solution level	True $\mu_{pi} - \mu_{Li}$	
Uncorrected Kroll vs. Corrected Kroll	5	0.05	1	-.23905	
			2	0.11952	
			3	0.23905	
			4	0.11952	
			5	-.23905	
	7	0.05	1	-.28868	
			2	0.00000	
			3	0.17321	
			4	0.23094	
			5	0.17321	
			6	.00000	
			7	-.28868	
	Estimation Method vs. TOST	5	0.01656	1	-.20000
				2	.10000
3				0.20000	
4				0.10000	
5				-.20000	
7		0.01017	1	-.20000	
			2	0.00000	
			3	0.12000	
			4	0.16000	
			5	0.12000	
			6	.00000	
			7	-.20000	

The reason for the extremely high empirical size of the uncorrected Kroll method is from an incorrect formulation of the hypothesis for proving the linearity of the analytical methods. The type I error with respect to proving the linearity is the error that the analytical method is claimed to be linear but in fact it is not. Therefore,

**Table 2** Specifications of parameters for power

No. of solution levels	True ADL	Solution level	True $\mu_{pi} - \mu_{Li}$
5	0.00151	1	-0.02
		2	0.01
		3	0.02
		4	0.01
		5	-0.02
7	0.00494	1	-0.10
		2	0.00
		3	0.06
		4	0.08
		5	0.06
		6	0.00
		7	-0.10

Table 3 Results of empirical sizes

Type of comparisons	No. of sol.	No. of rep.	SD	Kroll's Method		EP6-A	TOST	
				Uncorr.	Corr.			
Kroll – uncorr. vs. corr.	5	2	0.1	0.9232	0.0662	0.1590	0.0056	
			0.2	0.9292	0.0780	0.3030	0.0166	
		3	0.1	0.9328	0.0636	0.1050	0.0010	
			0.2	0.9332	0.0646	0.2632	0.0114	
		4	0.1	0.9420	0.0624	0.0768	0.0006	
			0.2	0.9328	0.0608	0.2296	0.0074	
		7	2	0.1	0.9328	0.0622	0.0110	0.0000
				0.2	0.9332	0.0606	0.1328	0.0028
	3		0.1	0.9418	0.0604	0.0024	0.0000	
			0.2	0.9458	0.0576	0.0754	0.0012	
	4		0.1	0.9438	0.0516	0.0002	0.0000	
			0.2	0.9424	0.0586	0.0544	0.0008	
	Estimation vs. TOST.	5	2	0.1	1.0000	1.0000	0.4984	0.0522
				0.2	0.9998	1.0000	0.5050	0.0564
3			0.1	1.0000	1.0000	0.4930	0.0440	
			0.2	1.0000	1.0000	0.5050	0.0486	
4			0.1	1.0000	1.0000	0.5048	0.0490	
			0.2	1.0000	1.0000	0.5066	0.0560	
7			2	0.1	0.9998	0.9998	0.4972	0.0512
				0.2	1.0000	1.0000	0.5024	0.0484
		3	0.1	0.9998	0.9998	0.4978	0.0478	
			0.2	1.0000	1.0000	0.4946	0.0504	
		4	0.1	1.0000	1.0000	0.5044	0.0504	
			0.2	1.0000	1.0000	0.5066	0.0494	

Sol.: solution; Rep.: replications; Uncorr.: uncorrected; Corr.: correction.

the empirical size of the uncorrected Kroll method at the 5% nominal level should be close to 95%. On the contrary, the empirical size of the corrected Kroll method should be close to 5% at the 5% nominal level. However, one needs to estimate the noncentral parameters for noncentral  $\chi^2$  distribution of the observed ADL. In addition, the critical value in Eq. (7) also contains an estimator,  $\bar{X}$ . Therefore, both the uncorrected and corrected Kroll methods ignore the variability of the estimators in the noncentral parameters and critical value. As a result, although the empirical size of the corrected Kroll method is close to 0.05, it is still inflated above the nominal level of 0.05. The empirical sizes of the estimation method and TOST procedure for the same specifications are also provided in Table 3. From Table 1, when the true ADL is 0.05,  $\mu_{Pi} - \mu_{Ci}$  at some solutions is either greater than 0.2 or smaller  $-0.2$ . It follows that all empirical sizes of the TOST procedure are less than 0.02. However, on the contrary, the empirical size of estimation method suggested in the approved CLSI guideline EP6-A can reach as high as 0.30 even when the differences in means between the best-fitted curve and the linear regression equation are outside the margin of  $(-0.2, 0.2)$  at three of the five solutions.

For the comparison between the estimation method in the approved CLSI guideline EP6-A and TOST procedure, the empirical sizes of the TOST procedure ranges from 0.0440 to 0.0564. Only 8.33% of the empirical sizes (1/12) are not

**Table 4** Results of empirical powers

No. of sol.	No. of rep.	SD	Kroll's method		EP6-A	TOST
			Uncorr.	Corr.		
5	2	0.1	1.0000	1.0000	0.9954	0.7616
		0.2	1.0000	1.0000	0.9052	0.2976
	3	0.1	1.0000	1.0000	0.9998	0.9232
		0.2	1.0000	1.0000	0.9470	0.4452
	4	0.1	1.0000	1.0000	0.9998	0.9754
		0.2	1.0000	1.0000	0.9664	0.5518
7	2	0.1	1.0000	1.0000	0.9954	0.7754
		0.2	1.0000	1.0000	0.9014	0.3168
	3	0.1	0.9998	0.9998	0.9994	0.9164
		0.2	1.0000	1.0000	0.9450	0.4468
	4	0.1	1.0000	1.0000	1.0000	0.9704
		0.2	1.0000	1.0000	0.9660	0.5570

Sol.: solution; Rep.: replications; Uncorr.: uncorrected; Corr.: correction.

contained within (0.04395, 0.05604). The one outside (0.04395, 0.05604) has the empirical size of 0.0564, which is just 0.0036 above 0.05604. However, the range of the empirical sizes of the estimation method is from 0.4930 to 0.5066. Recall that the estimation method suggested in the approved CLSI guideline EP6-5A ignores the variation of the estimates of  $\mu_{P_i} - \mu_{C_i}$ . When  $\mu_{P_i} - \mu_{C_i}$  is equal to either 0.2 or  $-0.2$  at some solutions, and then under the normal assumption, the size should be equal to 0.5 as confirmed by the empirical sizes of the simulation.

Table 4 presents the results of the empirical powers. For the simulation, the true ADL is specified as 0.00151 or 0.00494 when the number of solutions is 5 or 7, respectively. Therefore, with an allowable margin of 5%, 91.67% of the empirical powers of the uncorrected and corrected Kroll methods reach 1. On the other hand, the empirical powers of the estimation method and TOST procedures are smaller than those of the uncorrected and corrected Kroll methods. In addition, the empirical powers of the estimation method suggested in the approved CLSI guideline EP5-A and TOST procedures increase as the number of replicates increases or the standard deviation decreases. The results in Table 4 show that the empirical power of the estimation method is greater than that of the TOST procedure. However, from Table 3, the uncorrected and corrected Kroll methods, and the estimation procedure fail to control the size at the nominal level. Therefore, the advantage of power by these methods comes at the expense of inflation of the type I error rate. From the results of the simulation in Table 4, the power of the TOST procedure is greater than 0.9 when the standard deviation is 0.1 and number of replicates is at least 3.

## NUMERICAL EXAMPLES

We consider an experiment for the evaluation of linearity of a new analytical procedure for determination of  $\beta$ -HCG ( $\beta$ -Human Chorionic Gonadotropic, mIU/mL). The design consists of 5 dilutions with two replicates at each dilution

**Table 5** Measurement of  $\beta$ -HCG

Dilution	Replicate 1	Replicate 2
1	1.00	0.99
2	1.60	1.59
3	2.50	2.60
4	4.36	4.39
5	5.10	5.00

of concentration. Table 5 presents a set of hypothetic measurements under the described design. For the purpose of the illustration, the allowable margin of percent bound for ADL is set as 0.05 for uncorrected and corrected Kroll methods. On the other hand, the allowable limit of  $|\mu_{Pi} - \mu_{Li}|$  is set as 0.4 for the estimation method and for the TOST procedure. Table 6 provides the results of regression analyses for the linear, quadratic, and cubic linear regression models. The results of the regression analyses presented in Table 6 demonstrate that all estimates of the regression coefficients of the cubic model are significantly different from 0 at the 5% level ( $t_{0.025,6} = 2.4469$ ). In addition, the standard error of the residuals from the estimated cubic regression equation is 0.1799, which is at least 40% smaller than those from the linear or the quadratic models. Furthermore, the coefficient of determination,  $R^2$ , is also above 0.99. As a result, the cubic model is the best-fitted model among the three models recommended by the approved CLSI guideline EP6-A. Figure 1 presents the fitted the cubic, linear regression equations and the means at each of the five dilutions. It clearly shows that the relationship between the dilutions of concentrations and the analytical results is nonlinear and the cubic model is a better fit than the simple linear regression model.

Table 7 gives the predicted means from the cubic and linear regression models at each of the five dilutions as well as their corresponding differences, while Table 8 provides summarized results of linearity by the four methods. From these differences and observed mean concentrations, the observed ADL yields a value of 0.0842. With respect to the hypothesis in Eq. (5) and a margin of percent bound of 5%, the critical value in Eq. (7) is 0.0851, which is greater than the observed ADL of 0.0842.

**Table 6** Summary of results of regression analyses

Order	Coefficient	Value	SE	$t$ -test	Std err $S_{y-x}$	Degrees freedom
Linear	$\alpha'$	-0.354	0.234	-1.51	0.3154	8
	$\beta'_1$	1.089	0.071	15.44		
Quadratic	$\alpha''$	0.156	0.461	0.34	0.3041	7
	$\beta''_1$	0.652	0.351	1.85		
	$\beta''_2$	0.073	0.058	1.27		
Cubic	$\alpha'''$	2.263	0.626	3.62	0.1799	6
	$\beta'''_1$	-2.308	0.818	-2.82		
	$\beta'''_2$	1.202	0.304	3.96		
	$\beta'''_3$	-0.125	0.034	-3.74		

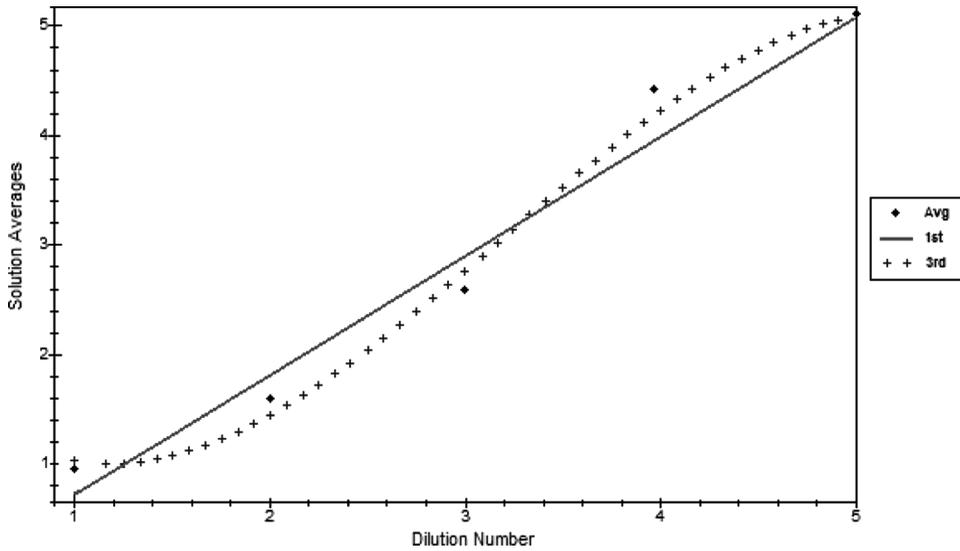


Figure 1 Regression curves for cubic vs. linear models of  $\beta$ -HCG.

According to the decision rule of the uncorrected Kroll method, the analytical method can conclude linearity at the 5% significance level. However, it should be noted that for this example, even though the observed ADL of 0.0842 is already greater than the allowable percent bound of 0.05, the linearity of the analytical method still can be claimed by the uncorrected Kroll method. On the other hand, with respect to the hypothesis in Eq. (10) for the corrected Kroll method, the critical value with an allowable margin of 5% in Eq. (11) is 0.0237. Since the observed ADL of 0.0842 is greater than 0.0237, we cannot reject the null hypothesis and cannot conclude linearity of the analytical method at the 5% significance level. Unlike the uncorrected Kroll method, the conclusion of the corrected Kroll method is consistent with the evidence, for which the observed ADL is 0.0842, which is greater than the allowable percent bound of 0.05.

With respect to the estimation method suggested in the approved CLSI guideline EP6-A, the observed differences in the predicted means between the cubic and linear regression models at all dilutions are within the allowable margin of  $\pm 0.4$ . As a result, the linearity is claimed by the estimation method. On the other hand, the results of the TOST procedure show that the 95% confidence intervals for

Table 7 Mean differences between the best-fitted curve and simple linear regression equation

Result mean	Predicted (linear)	Predicted (cubic)	Difference	% Difference
0.995	0.735	1.031	0.296	28.7
1.595	1.824	1.450	-0.374	25.8
2.550	2.913	2.767	-0.146	5.3
4.375	4.002	4.230	0.228	5.4
5.050	5.091	5.086	-0.005	0.1

**Table 8** Results of the linearity by four different methods

Sample ADL	Kroll		TOST				Estimation method				
	Uncorrected		Corrected		Dil. (i)	90% C.I. of $\mu_{Pi} - \mu_{Li}$	Result		Result		
	Critical value	Result	Critical value	Result			Ind.	Overall	$\hat{Y}_{Pi} - \hat{Y}_{Li}$	Ind.	Overall
0.0842	0.0851	L	0.0237	NL	1	(0.064, 0.529)	NL	NL	0.296	L	L
					2	(-0.524, -0.223)	NL	NL	-0.374	L	L
					3	(-0.318, 0.027)	L	L	-0.146	L	L
					4	(0.078, 0.379)	L	L	0.228	L	L
					5	(-0.237, 0.228)	L	L	-0.005	L	L

Dil.: Dilution level; Ind.: Individual dilutions; Overall: Overall conclusion; NL: Conclusion of nonlinearity at the 5% nominal level; L: Conclusion of linearity at the 5% nominal level.

$\mu_{P_i} - \mu_{C_i}$  at the first two dilutions are not contained within  $(-0.4, 0.4)$ . With respect to hypotheses in Eq. (8), the analytical method cannot conclude linearity at the 5% significance level. Because the estimation method completely ignores the variability in the observed differences in the predicted means, its conclusion is made without any statement of the probability of type I error. However, in fact as demonstrated by the simulation, the probability of a type I error in the estimation method far exceeds its nominal significance level.

## DISCUSSION

Linearity is one of the most important characteristics for evaluation of accuracy and precision in assay validation. Various methods have been proposed for evaluation of linearity. With respect to the disaggregate criterion, the estimation method suggested by the approved CLSI guideline ignores the variation in the estimates of the differences in the predicted means and is not a formal statistical inference procedure. On the other hand, the procedure based on the aggregate criterion of ADL proposed by Kroll et al. (2000) incorrectly formulated the hypothesis for proving linearity as the null hypothesis. As a result, the uncorrected Kroll method cannot control the type I error in decision making regarding conclusions of linearity. As a result, we proposed the TOST procedure for the disaggregate criterion and the corrected Kroll method for the aggregate criterion based on ADL by formulating the hypothesis for proving linearity as the alternative hypothesis. Simulation results and the preceding numerical examples demonstrate that the proposed TOST and the correct Kroll method can not only adequately control the type I error rate but also reach a conclusion consistent with the data.

Hypotheses in Eq. (1) require that  $|\mu_{P_i} - \mu_{L_i}| < \delta$ , for all  $i = 1, \dots, L$ . Therefore, the statement that  $|\mu_{P_i} - \mu_{L_i}| < \delta$ , for all  $i = 1, \dots, L$  is in fact equivalent to that  $\max_{1 \leq i \leq L} |\mu_{P_i} - \mu_{L_i}| < \delta$ . Therefore the hypothesis for proving that the difference in the predicted means between the best-fitted and linear regressions is within the pre-specified allowable limit at every concentration level can also be reformulated as

$$H_0 : \max_{1 \leq i \leq L} |\mu_{P_i} - \mu_{L_i}| \geq \delta_0 \quad \text{vs.} \quad H_a : \max_{1 \leq i \leq L} |\mu_{P_i} - \mu_{L_i}| < \delta_0. \quad (12)$$

However, the development of a testing procedure for hypotheses in Eq. (12) requires further research.

Since the TOST procedure is constructed based on a disaggregate criterion, which requires that all differences in the predicted means between the best-fitting and linear models be within pre-specified allowable limits, the method is more conservative than the corrected Kroll method, which is based on an aggregate criterion and only requires ADL, a function of standardized sum-of-squares of the differences in the predicted means between the best-fitted and linear models to be controlled within the pre-specified allowable percent bound. However, as previously mentioned, the inference based on ADL involves the estimation of the unknown noncentrality parameter and the average population mean concentration. When these estimates are assumed fixed constants for the inference based on ADL, the simulation study shows that the empirical size can be inflated up to 0.078 at the

0.05 significance level. The disaggregate criterion required by hypotheses in Eq. (2) or Eq. (8) can be considered as a hypothesis for the model-by-dilution interaction. However, the hypothesis is not to detect the existence of the model-by-dilution interaction but rather to verify whether the model-by-dilution interaction is within some pre-specified allowable upper limit. Therefore one of the possible hypotheses can be formulated as follows:

$$H_0 : \sum_{i=1}^L (\mu_{Pi} - \mu_{Li})^2 \geq L\delta_0^2 \quad \text{vs.} \quad H_0 : \sum_{i=1}^L (\mu_{Pi} - \mu_{Li})^2 < L\delta_0^2. \quad (13)$$

This hypothesis is very similar to the hypothesis of the corrected Kroll method based on ADL. However, the hypothesis in Eq. (13) avoids estimating the unknown overall concentration mean, which is only for the purpose of scaling. Preliminary research suggests the procedure based on F statistics may be better than that based on  $\chi^2$  statistics. However, further research and evaluation is needed.

## ACKNOWLEDGMENT

We would like to thank the referees for their careful, thoughtful, and thorough reviews, which greatly improve the contents and presentation of our work. This research is partially supported by the Taiwan National Science Council Grant: NSC 95 2118-M-002-007-MY2 to Jen-pei Liu.

## REFERENCES

- Chow, S. C., Liu, J. P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker, Inc.
- Clinical Laboratory Standard Institute. (2003). *EP6-A. Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach, Approved Guideline*. Wayne, PA: Clinical Laboratory Standard Institute.
- International Conference on Harmonization. (1995). Tripartite guideline Q2A: *Test on validation of analytical procedures*. Center for Drug Evaluation and Research, The U.S. Food and Drug Administration, Rockville, MD, USA.
- International Conference on Harmonization. (1996). Tripartite guideline Q2B: *Validation of analytical procedures: methodology*. Center for Drug Evaluation and Research, The U.S. Food and Drug Administration, Rockville, MD, USA.
- Kroll, M. H., Præstgaard, J., Michaliszyn, E., Styer, P. E. (2000). Evaluation of the extent of nonlinearity in reportable range studies. *Arch Pathol Lab Med* 124:1331–1338.
- Schofield, T. L. (2003a). Assay development. In: Chow, S. C., ed., *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker, Inc., pp. 55–62.
- Schofield, T. L. (2003b). Assay validation. In: Chow, S. C., ed., *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker, Inc., pp. 63–71.
- Zhong, B., Qiu, C., Tang D. (2003). Validation of quantitative and qualitative assays. In: Chow, S. C., ed., *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker, Inc., pp. 1–14.

Copyright of Journal of Biopharmaceutical Statistics is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.