Research Article

JOURNAL OF CHEMOMETRICS

Received: 28 February 2008, Revised: 21 June 2008, Accepted: 20 August 2008, Published online in Wiley InterScience: 6 October 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1194

Statistical methods for evaluating the linearity in assay validation†,‡

Eric Hsieh[a], Chin-fu Hsiao[b] and Jen-pei Liu[a,b]*

One of the most important characteristics for evaluation of the accuracy in assay validation is the linearity. Kroll, *et al*. [1] proposed a method based on the average deviation from linearity (ADL) to evaluate the linearity. Hsieh and Liu [2] suggested that hypothesis for proving the linearity be formulated as the alternative hypothesis and proposed the corrected Kroll's method. However, the issue concerning the variability in estimation of the non-centrality parameter is still unresolved. Consequently, the type I error rates may still be inflated for the corrected Kroll's method. To overcome this issue, we propose the sum of squares of deviations from linearity (SSDL) as an alternative metric for evaluation of linearity. Based on SSDL, we applied the method of generalized pivotal quantities (GPQ) for the inference of evaluation of linearity. The simulation studies were conducted to empirically investigate the size and power between current and proposed methods. The simulation results show that the proposed GPQ method not only adequately control size but also provide sufficient power than other methods. A numeric example illustrates the proposed methods. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: linearity; allowable limit; quantitative analytical laboratory methods; bootstrap; general pivotal quantities

1. INTRODUCTION

In validation of quantitative analytical laboratory procedures, one of the most important characteristics of the accuracy is the linearity. The ICH Q2A guideline [3], defines the linearity of an analytical method 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. The objective for evaluation of linearity is to validate existence of a mathematically verified straight-line relationship between the observed values and the true concentrations or activities of the analyte. Linearity represents the simplest mathematical relationship and it also permits simple and easy interpolations of results for clinical practitioners. The approved Clinical Laboratory Standard Institute (CLSI) guideline EP6-A [4] recommends that at least five solutions of different concentration levels across the anticipated range be included in an experiment for evaluation of linearity. At each concentration level, two to four replicates should be run. With respect to EP6-A, if the difference between the best-fit nonlinear polynomial curve and simple linear regression equation at each concentration is smaller than some pre-defined allowable bias, the linearity then can be claimed. On the other hand, Kroll, *et al*. [1] proposes a statistical test procedure based on the average deviation from linearity (ADL) which is defined as the square root of the average squared distances between the fitted concentrations of the best fit polynomial curve and the simple regression equation at each solution level, standardized by mean concentration. The linearity is concluded at the α nominal level if the observed value of the ADL is smaller than the upper α quantile of the sampling distribution of the observed ADL. However, the sampling distribution of the observed ADL is a function of a non-central chi-square distribution. It follows that the Kroll's method suggests using the estimate of the unknown non-centrality parameter as the true parameter. Hence, the variability associated with the estimated non-centrality parameter is completely ignored in the Kroll's procedure.

On the other hand, Hsieh and Liu [2] pointed out that the Kroll's method based on the ADL is derived from an improper formulation of hypotheses and suggested that the hypothesis of proving the linearity should be formulated as the alternative hypothesis. They termed their proposed procedure as the corrected Kroll's method. However, the issue concerning the variability in estimation of the non-centrality parameter is still unresolved for the corrected Kroll's method. Consequently, the type I error rates may still be inflated for the corrected Kroll's method. As mentioned before, the approved CLSI EP6-A recommends that for proving the linearity, the deviations from linearity, defined as the difference between the best-fit nonlinear polynomial curve and simple linear regression equation, be smaller than some pre-defined allowable bias, say δ_0, at all concentrations. Therefore, we propose the sum of squares of

* *Correspondence to: J. P. Liu, Division of Biometry, Department of Agronomy, National Taiwan University, 1, Section 4, Roosevelt Road, Taipei, Taiwan. E-mail: jpliu@ntu.edu.tw*

a *E. Hsieh, J. P. Liu*
Division of Biometry, Institute of Agronomy, National Taiwan University, Taipei, Taiwan

b *C. F. Hsiao, J. P. Liu*
Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan, Taiwan

† *Eric Hsieh and Chin-fu Hsiao contribute equally to this work.*

‡ *The views expressed in this article are personal opinions of the authors and may not necessarily represent the position of National Taiwan University, and the National Health Research Institutes, Taiwan.*

deviations from linearity (SSDL) as an alternative metric for evaluation of the linearity in assay validation. Tsui and Weerahandi [5], and Weerahandi [6] propose the generalized confidence interval (GCI) based on the generalized pivotal quantity (GPQ) for the exact statistical inference. It has been successfully applied to various areas including population and individual bioequivalence [7], tolerance intervals for quality control [8,9], and the area under the receiver operating characteristic (ROC) curve [10,11]. As a result, we propose to apply the concept of GCI for validation of quantitative analytical laboratory procedures. In addition, we also apply the bootstrap method [12] to linearity validation. In Section 2, we review experiment designs for evaluation of linearity recommended by the approved CLSI guideline EP6-A [4] as well as the corrected Kroll's methods proposed by Hsieh and Liu [2]. The impact of ignoring the variability of estimated non-centrality parameters on the size of the corrected Kroll's method is highlighted. In addition, the proposed exact methods based on the GCI and bootstrap method using SSDL are provided in Section 2. In Section 3, the results of a simulation study for investigation of empirical size and power of various methods are reported. Numeric examples are employed to illustrate and compare the current and proposed methods. Discussion and Conclusion are also provided in this section.

2. MATHEMATICAL SETTING

2.1. Experiment design

The experiments for linearity assessment, as recommended in the approved CLSI guideline EP6-A [4], require that at least five solutions of different concentrations be run at least in duplicates. The following linear, quadratic, and cubic models are also considered in the guideline for fitting the data obtained from the experiment

$$\begin{aligned}&\text{Linear } \mu_{\text{Li}} = \alpha' + \beta_1' \text{X}_i \\ &\text{Quadratic } \mu_{\text{Qi}} = \alpha'' + \beta_1'' \text{X}_i + \beta_2'' \text{X}_i^2 \\ &\text{or} \\ &\text{Cubic } \mu_{\text{Ci}} = \alpha''' + \beta_1''' \text{X}_i + \beta_2''' \text{X}_i^2 + \beta_3''' \text{X}_i^3\end{aligned} \tag{1}$$

where μ_{Li}, μ_{Qi}, and μ_{Ci} are the predicted mean of the respective models and α', α'', α'''; β_1', β_1'', β_1'''; β_2'', β_2''', and β_3''' are the intercepts, regression coefficients for the corresponding models in (1). The approved CLSI guideline EP6-A [4] also suggests that the best-fitted model be used in the linearity assessment. The best-fit model defined in the EP6-A [4] is the model such that the lack-of-fit of the model is not statistically significant and the repeatability meets the manufacturer's claim. Furthermore, we also assume that the random error is approximately constant in the range of concentrations employed by the experiment. The following two conditions for claiming the linearity of an analytical procedure are recommended by the approved CLSI guideline EP6-A [4]:

(a) If the best-fitted model is the linear model over the same range of concentrations employed in the experiment.
(b) If the best-fitted model is not linear, the magnitude of deviations from the linearity at each concentration is within some pre-specified allowable limit of δ_0.

2.2. Current testing procedures

Denote μ_{Pi} as the predicted mean of the best-fitted model, where μ_{Pi} can be either μ_{Qi} or μ_{Ci}. The deviation from linearity at each concentration level is defined as the difference in the predicted means between the best-fit nonlinear model and linear model $\mu_{\text{Pi}}-\mu_{\text{Li}}$. The average deviation from linearity (ADL) suggested by Kroll, *et al.* [1] for assessment of linearity is defined as

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

where μ is the population mean concentration for all solutions of the assay.

Therefore, Kroll, *et al.* [1] suggest the following hypothesis for evaluation of linearity based on the ADL

$$\text{H}_0 : \theta \le \theta_0 \text{ versus } \text{H}_a : \theta > \theta_0 \tag{3}$$

where θ_0 is the maximum allowable per cent bound and is recommended as 0.05 by Kroll, *et al.* [1].

However, Hsieh and Liu [2] pointed out that the formulation of hypothesis given in Equation (3) is not appropriate for linearity validation and they suggested that the hypothesis of proving the linearity formulated as the alternative hypothesis as follows:

$$\text{H}_0 : \theta \ge \theta_0 \text{ versus } H_{\text{a}} : \theta < \theta_0 \tag{4}$$

Let Y_{ij} be the test result of replicate j at concentration X_i, where $j = 1,\ldots, J$; $i = 1,\ldots, L$, and $\hat{Y}_{Pi}$ and $\hat{Y}_{Li}$ denote the least squares (LS) estimators of the predicted mean of the best-fit and linear models, respectively, where

$$\hat{Y}_{Li} = a' + b_1' \text{X}_i$$

and

$$\hat{Y}_{\text{Pi}} = \begin{cases} \text{a}'' + \text{b}''_1 \text{X}_i + \text{b}''_2 \text{X}_i^2, & \text{if quadratic} \\ \text{a}''' + \text{b}'''_1 \text{X}_i + \text{b}'''_2 \text{X}_i^2 + \text{b}'''_3 \text{X}_i^3, & \text{if cubic} \end{cases}$$

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).

An estimator of ADL is given as

$$\hat{\theta} = \frac{\sqrt{\sum_{i=1}^{\text{L}} \left(\hat{Y}_{\text{Pi}} - \hat{Y}_{\text{Li}}\right)^2 / \text{L}}}{\overline{X}}$$

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

Note that $\sum_{i=1}^{\text{L}} (\hat{Y}_{\text{Pi}} - \hat{Y}_{\text{Li}})^2$ follows a non-central χ^2 distribution with degrees of freedom d−1, and non-centrality parameter $LJ\theta_0^2/(\sigma/\mu)^2$ [1,13], where d is the degrees of freedom for regression of the best-fitted model, μ is the population mean concentration for all solutions of the assay and σ^2 is the variance of residuals under the best-fitted model. It follows that the null hypothesis in Equation (4) is rejected and the linearity of an analytical procedure is concluded at the 5% significance level if

$$\hat{\theta} < \frac{\sigma}{\mu} \sqrt{\frac{\text{q}_{0.05}}{\text{LR}}} \tag{5}$$

where $q_{0.05}$ is the 5th percentile of a non-central χ^2 distribution with degrees of freedom d−1 and non-centrality parameter $LJ\theta_0^2/(\sigma/\overline{X})^2$.

Note that the critical value in (5) contains the unknown parameters μ, σ, and the non-centrality parameter. One way to resolve this issue is to estimate μ by $\overline{X}$, the observed mean concentration for all solutions of the assay, σ by the square root of residual mean square obtained the best-fitted model, $\hat{\sigma}_e$, and $LJ\theta_0^2/(\hat{\sigma}_e/\overline{X})^2$ for the non-centrality parameter. Although the corrected Kroll's method has the appropriate formulation of hypotheses for proving the assay linearity, it completely ignores the variability in estimation of the non-centrality parameter. Consequently, probability of incorrectly concluding linearity when the assay is in fact nonlinear is inflated.

2.3. Proposed testing procedures

2.3.1. Statistical hypothesis

According to the approved CLSI guideline EP6-A [4], the linearity of the proposed analytical method can be concluded if the deviation from linearity is smaller than some pre-specified limit δ_0 at all concentrations

$$|\mu_{Pi} - \mu_{Li}| < \delta_0, \quad \text{for } i = 1, \ldots, L$$

As a result, a natural aggregate metric for assessment of assay linearity is the SSDL defined as

$$\text{SSDL} = \sum_{i=1}^{L} (\mu_{Pi} - \mu_{Li})^2 \tag{6}$$

It follows that the hypotheses for proving the assay linearity can be formulated based on SSDL as follows:

$$H_0 : \sum_{i=1}^{L} (\mu_{Pi} - \mu_{Li})^2 \geq L\delta_0^2 \text{ versus } H_0 : \sum_{i=1}^{L} (\mu_{Pi} - \mu_{Li})^2 < L\delta_0^2 \tag{7}$$

or equivalently

$$H_0 : \sum_{i=1}^{L} (\mu_{Pi} - \mu_{Li})^2/L \geq \delta_0^2 \text{ versus } H_0 : \sum_{i=1}^{L} (\mu_{Pi} - \mu_{Li})^2/L < \delta_0^2$$

As a result, applications of the GCI and the bootstrap method to hypothesis (7) of evaluation of assay validation are provided in the following subsequent subsections.

2.3.2. Generalized pivotal quantity approach

2.3.2.1. Generalized confidence interval

2.3.2.1.1. Generalized pivotal quantity for SSDL Suppose that V is a random variable whose distribution depends on a vector of unknown parameters $\zeta = (\theta, \boldsymbol{\eta})$, where θ is a parameter of interest and $\boldsymbol{\eta}$ is a vector of nuisance parameter. Let $\mathbf{V}$ be a random sample from V and $\mathbf{v}$ be the observed value of $\mathbf{V}$. Also let $\boldsymbol{R} = \boldsymbol{R}(\mathbf{V}; \mathbf{v}, \zeta)$ be a function of $\mathbf{V}$, $\mathbf{v}$ and ζ. The random quantity $\boldsymbol{R}$ is said to be a GPQ if it satisfies the following two conditions:

(a) The distribution of $\boldsymbol{R}$ does not depend on any unknown parameter.

(b) The observed value of $\boldsymbol{R}$, say $r = \boldsymbol{R}(\mathbf{V}; \mathbf{v}, \zeta)$, is free of the vector of nuisance parameters $\boldsymbol{\eta}$. In other words, the value of $\boldsymbol{R}$ at $\mathbf{V} = \mathbf{v}$ should be a function only of $(\mathbf{v}, \theta)$.

Specifically, if the observed quantity $r = \theta$, then the GPQ is called the fiducial generalized pivotal quantity (FGPQ) and GCI based on FGPQ are proven to have asymptotically correct frequent coverage probability in Hanning *et al.* [14]. In consequence, an upper $100(1-\alpha)$th percentile GCI for θ is given by $(0, R_{1-\alpha})$, where $R_{1-\alpha}$ are the $100(1-\alpha)$th percentile of the distribution of $\boldsymbol{R}$. The percentile of $\boldsymbol{R}$ can be estimated using Monte-Carlo algorithms.

For the regression models in Equation (1), define

$$\begin{aligned}
&\mathbf{Y} = \text{the LJ} \times 1 \text{ vector of observations,}\\
&\mathbf{X}_L = (\mathbf{1}, \mathbf{X}),\\
&\mathbf{X}_P = \begin{cases} (\mathbf{1}, \mathbf{X}, \mathbf{X}_2), \text{ if the best-fitted model is quadratic, and} \\ (\mathbf{1}, \mathbf{X}, \mathbf{X}_2, \mathbf{X}_3), \text{ if the best-fitted model is cubic,} \end{cases}\\
&\boldsymbol{\mu}_P = \text{the LJ} \times 1 \text{ predicted mean vector of best-fitted polynomial model, and}\\
&\boldsymbol{\mu}_L = \text{the LJ} \times 1 \text{ predicted mean vector of linear model}
\end{aligned}$$

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

Let $\mathbf{W} = (\mathbf{W}_P - \mathbf{W}_L)$, $\mathbf{W}_P$ and $\mathbf{W}_L$ be 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 $\mathbf{W}_L = \mathbf{X}_L(\mathbf{X}'_L\mathbf{X}_L)^{-1}\mathbf{X}'_L$. $\hat{\mathbf{Y}}_P = \mathbf{W}_P\mathbf{Y}$ and $\hat{\mathbf{Y}}_L = \mathbf{W}_L\mathbf{Y}$ are then the LS estimators of the predicted mean vectors of the best-fit and linear models. As a result, the unbiased and sufficient estimator of $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$ and its covariance matrix, $\boldsymbol{\Sigma}$, are given as respectively

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}_P - \hat{\boldsymbol{\mu}}_L = \hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L = \mathbf{WY}\\
&\text{Cov}\left(\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L\right) = S^2\mathbf{WW}'
\end{aligned} \tag{8}$$

where S^2 is the residual mean square obtained from the best-fitted polynomial model with degree of freedom $LJ-d-1$. Under the assumption that random errors in model (1) are identically and independently distributed as normal distribution with mean of zero and variance of σ^2, $\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L$ is distributed as a multi-normal distribution with mean and variance $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$ and $\boldsymbol{\Sigma}$ which is equal to $\sigma^2\mathbf{WW}'$.

It is easy to verify that the estimators $\mathbf{WY}$ and S^2 are associated with pivotal quantities $\mathbf{Z}$ and U which are independent with the following distributions

$$\begin{aligned}
&\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}[\mathbf{WY} - (\boldsymbol{\mu}_P - \boldsymbol{\mu}_L)] \sim \mathbf{N}_{LJ}(\mathbf{0}, \mathbf{I})\\
&U = \frac{(LJ-d-1)S^2}{\sigma^2} \sim \chi^2_{LJ-d-1}.
\end{aligned} \tag{9}$$

Here matrix $\boldsymbol{\Lambda}^{1/2}$ denotes the positive definite square root of a positive definite matrix $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^{-1/2} = (\boldsymbol{\Lambda}^{1/2})^{-1}$. Then $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$ can be expressed as

$$\begin{aligned}
\boldsymbol{\mu}_P - \boldsymbol{\mu}_L &= \mathbf{WY} - \boldsymbol{\Sigma}^{1/2}\mathbf{Z}\\
&= \mathbf{WY} - \left(\sigma^2\mathbf{WW}'\right)^{1/2}\mathbf{Z}\\
&= \mathbf{WY} - \left(\frac{(LJ-d-1)S^2}{U}\mathbf{WW}'\right)^{1/2}\mathbf{Z}
\end{aligned} \tag{10}$$

Let $\mathbf{y}$ and S^2 be the observed values of $\mathbf{Y}$ and S^2, respectively, the GPQ for $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$ is given by

$$\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L} = \mathbf{W}\mathbf{y} - \left(\frac{(LJ - d - 1)s^2}{U}\mathbf{W}\mathbf{W}'\right)^{1/2}\mathbf{Z} \quad (11)$$

$$= \mathbf{W}\mathbf{y} - \left(\frac{s^2\sigma^2}{S^2}\mathbf{W}\mathbf{W}'\right)^{1/2}\boldsymbol{\Sigma}^{-1/2}[\mathbf{W}\mathbf{Y} - (\boldsymbol{\mu}_P - \boldsymbol{\mu}_L)] \quad (12)$$

From Equation (11), $\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L}$ has distribution that is free of parameters. When $\mathbf{Y}$ and S^2 are substituted by their observed values $\mathbf{y}$ and s^2 in Equation (12), then $\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L}$ turns out to be $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$. Hence, it fulfills the requirements of (a) and (b) and $\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L}$ is a GPQ for $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$. Moreover, since $\mathrm{SSDL} = (\boldsymbol{\mu}_P - \boldsymbol{\mu}_L)'(\boldsymbol{\mu}_P - \boldsymbol{\mu}_L)/J$, where J is the number of replicates of observation of $\mathbf{Y}$, a GPQ of SSDL can be obtained as

$$\boldsymbol{R}_{\mathrm{SSDL}} = \frac{1}{J}\left(\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L}\right)'\left(\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L}\right) \quad (13)$$

where $\boldsymbol{R}_{\boldsymbol{\mu}_P-\boldsymbol{\mu}_L}$ is defined as Equation (11).

2.3.2.1.2. Generalized confidence interval for SSDL An upper $100(1-\alpha)$th percentile GCI for SSDL can be obtained from the following Monte-Carlo algorithm:

Step 1: Choose a large simulation sample size, say $K = 10\,000$. For k equal to 1 through K, carry out the following two steps.
Step 2: Generate $\mathrm{LR} \times 1$ standard normal random vector $\mathbf{Z}$ and central χ^2 random variable U with degree of freedom $LJ-d-1$.
Step 3: For the realized values of $\mathbf{Y}$ and S^2, compute $R_{\mathrm{SSDL},k}$ defined in Equation (13).

The required upper $100(1-\alpha)$th percentiles of the distribution of GPQ for SSDL, which is also the upper $100(1-\alpha)$th generalized confidence limit for SSDL, is then estimated by the $100(1-\alpha)$th sample percentiles of the collection of $K = 10\,000$ realizations $R_{\mathrm{SSDL},1}, R_{\mathrm{SSDL},2} \ldots, R_{\mathrm{SSDL},10\,000}$.

2.3.2.2. Statistical testing procedure
The upper $100(1-\alpha)$% generalized confidence limit for SSDL based on GPQ can be used to test the statistical hypothesis in Equation (7) for linearity. The null hypothesis in Equation (7) is rejected and the linearity of a analytical method is concluded at the α significance level if the upper $100(1-\alpha)$% generalized confidence limit for SSDL is less than $L\delta_0^2$. It should be noted that the method of the generalize *p*-values in the presence of nuisance parameters may provide more information about the parameter of interest. [5]

2.3.3. Bootstrap approach

2.3.3.1. Bootstrapping confidence interval
The nonparametric bootstrapping for regression model requires a large number of replicates at each concentration for ensuring the accuracy of bootstrap estimators of regression parameters. However, the experiment recommended by approved guideline EP6-A [4] suggested that only two to four replicates are needed for assessing linearity. Therefore, the nonparametric bootstrap method may not be appropriate to be implemented for linearity evaluation. On the other hand, the parametric bootstrap which is used when the distributional family of the data is considered known. Since $\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L$ is known as distributed as a multi-normal distribution with mean and variance $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$ and $\sigma^2\mathbf{W}\mathbf{W}'$, we propose the following parametric bootstrap algorithm to obtain an upper $100(1-\alpha)$% confidence interval for SSDL can be given as follows:

1. Use the observed response $\mathbf{y}$ to estimate the parameters, $\boldsymbol{\mu}_P - \boldsymbol{\mu}_L$ and $\sigma^2\mathbf{W}\mathbf{W}'$ by $\mathbf{W}\mathbf{y}$ and $s^2\mathbf{W}\mathbf{W}'$, respectively, where s^2 is residual mean square obtained from the best-fitted polynomial model.
2. Select a large number of bootstrap samples, say $B = 3000$, from the multi-normal distribution $N_{LJ}(\mathbf{W}\mathbf{y}, s^2\mathbf{W}\mathbf{W}')$ which are denoted by $\left(\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L\right)^{*1}, \left(\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L\right)^{*2}, \ldots, \left(\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L\right)^{*B}$.
3. Calculate bootstrap replication $SSDL_k^* = \left(\left(\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L\right)^{*k}\right)'\left(\left(\hat{\mathbf{Y}}_P - \hat{\mathbf{Y}}_L\right)^{*k}\right)/J$ corresponding to each bootstrap sample, for $k = 1, 2, \ldots, B$.
4. The upper $100(1-\alpha)$% confidence limit is given as $\hat{\mathrm{SSDL}}_{du} = \hat{\mathrm{SSDL}}_{*(1+(1-\alpha)B)}$, where $\hat{\mathrm{SSDL}}_{*(1)} < \hat{\mathrm{SSDL}}_{*(2)} < \ldots < \hat{\mathrm{SSDL}}_{*(B)}$ are the order statistics of bootstrap statistics $\hat{\mathrm{SSDL}}_1^*, \hat{\mathrm{SSDL}}_2^*, \ldots, \hat{\mathrm{SSDL}}_B^*$.

2.3.3.2. Statistical testing
The linearity of an analytical method can be concluded at the α significance level if $\hat{\mathrm{SSDL}}_{du}$, the bootstrap upper $100(1-\alpha)$% confidence limit for SSDL is less than $L\delta_0^2$.

3. SIMULATION STUDY

We conducted a simulation study to compare the empirical sizes and powers of the corrected Kroll's, parametric bootstrap and GPQ methods. Following the specification of the experiment 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, mean concentration μ is assumed to be 4 and the allowable margin of linearity based on ADL, θ_0, is specified 0.05 as recommended by Kroll *et al.* [1]. From the relationship that $\mathrm{SSDL} = L(\mu\theta)^2$, it follows that the margin for SSDL for 5 and 7 concentrations are 0.2 and 0.28, respectively. In addition, standard deviation of normal random error is specified as 0.1 and 0.2. For each of 12 combinations, ten thousand (10 000) random samples are generated. For the 5% nominal significance level, a simulation study with 10 000 random samples implies that 95 per cent of the empirical sizes evaluated at the allowable margins will be within 0.0457 and 0.0543 if the proposed methods can adequately control the size at the nominal level of 0.05.

Table I presents the results of the empirical sizes. All of empirical sizes for the corrected Kroll's and parametric bootstrap methods are larger than 0.0543. This indicates that both methods inflate the size and are quite liberal in concluding the linearity of an analytical procedure. On the other hand, all of empirical sizes of the GPQ method are within the range and showed that it has a good ability for controlling the size at the nominal level.

It is introduced in section 2.2 that the poor performance for the corrected Kroll's method in controlling the size results from σ, one of the components of non-central parameters for non-central χ^2 distribution of the observed ADL being estimated by the square

Table I. Empirical sizes

No. of solution	No. of replicate	Standard deviation	Corrected Kroll	Parametric bootstrap	GPQ
5	2	0.1	0.0769	0.0764	0.0535
		0.2	0.0734	0.0741	0.0503
	3	0.1	0.0679	0.0677	0.0523
		0.2	0.0643	0.0651	0.0501
	4	0.1	0.0569	0.0569	0.0476
		0.2	0.0596	0.0594	0.0504
7	2	0.1	0.0670	0.0673	0.0532
		0.2	0.0671	0.0669	0.0532
	3	0.1	0.0573	0.0582	0.0502
		0.2	0.0557	0.0553	0.0476
	4	0.1	0.0563	0.0571	0.0506
		0.2	0.0595	0.0593	0.0529

root of residual mean square obtained from best-fitted polynomial model. Compared the results of the corrected Kroll's with parametric bootstrap methods, the empirical sizes are very close to each other. It may indicate that the latter method cannot control the size due to the same reason since σ used for generating the parametric bootstrap samples is also estimated by the residual mean square obtained from the best-fitted model. In contrast, since one requirement for GPQ is that $\boldsymbol{R}_{\mu_P-\mu_L}$ is free of nuisance parameter σ, the GPQ approach is the only method under study that can control the size at the nominal level.

Table II presents the results of the empirical powers. For the simulation, the true ADL is specified as 0.005 for both number of solutions of 5 and 7. The results in Table II also show that the empirical power increases as the numbers of replicates or concentrations increases. All these methods provide comparable powers. However, it can be seen that the empirical powers of the GPQ method are smaller than all other methods for all combinations of parameters with the smallest powers is 0.6953 when number of solution is 5, number of replicates is 2, and standard deviation of normal random error is 0.2. However, all empirical powers of the GPQ method for other combinations of parameters are still greater than 90%. In addition, from Table I, both corrected Kroll's and parametric bootstrap methods fail to control the size at the nominal level. Therefore, the advantage of power by these two methods comes at the expense of inflated type I error rates.

Figures 1 and 2 present the empirical powers when the standard deviations of normal random error are 0.1 and 0.2, respectively with number of solutions is 5, number of replicates is 3, and the true ADLs are ranged from 0 to 0.08. Figure 1 show that when standard deviation is 0.1, the empirical size at ADL = 0.05 for the corrected Kroll's, and parametric bootstrap methods are 0.0679 and 0.0677, respectively, while the empirical size of the GPQ method is 0.0521. It shows that the GPQ method can control the size better than the other methods at the nominal level. In addition, the powers reach 0 and 1 at ADL = 0.08 and 0.005, respectively for all methods. The power curves of the corrected Kroll's and parametric bootstrap methods are almost overlapped from ADL of 0.01 to 0.06. On the other hand, the power of the GPQ method is quite competitive to the corrected Kroll's method and the bootstrap approach although it is little lower than other methods. The similar results are observed in Figure 2 when standard deviation of normal random error is 0.2. The empirical sizes for the corrected Kroll's, parametric bootstrap and the GPQ methods at ADL = 0.05 are 0.0827, 0.0618, and 0.0494, respectively. In addition, the powers for all three methods when the standard deviation is 0.2 are lower than those when the standard deviation is 0.1.

Table II. Empirical powers with the true ADL = 0.005

No. of solution	No. of replicate	Standard deviation	Corrected Kroll	Parametric bootstrap	GPQ
5	2	0.1	1.0000	1.0000	0.9994
		0.2	0.8306	0.8185	0.6953
	3	0.1	1.0000	1.0000	1.0000
		0.2	0.9454	0.9451	0.9256
	4	0.1	1.0000	1.0000	1.0000
		0.2	0.9828	0.9831	0.9781
7	2	0.1	1.0000	1.0000	1.0000
		0.2	0.9327	0.9317	0.9078
	3	0.1	1.0000	1.0000	1.0000
		0.2	0.9901	0.9900	0.9873
	4	0.1	1.0000	1.0000	1.0000
		0.2	0.9980	0.9980	0.9972

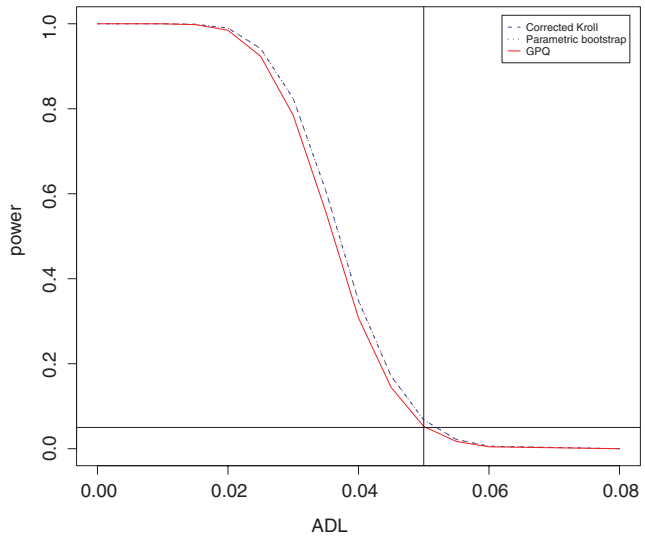

Figure 1. The empirical powers when standard deviation of normal random error is 0.1, number of solutions is 5, and number of replicates is 3. This figure is available in colour online at www.interscience.wiley.com/journal/cem

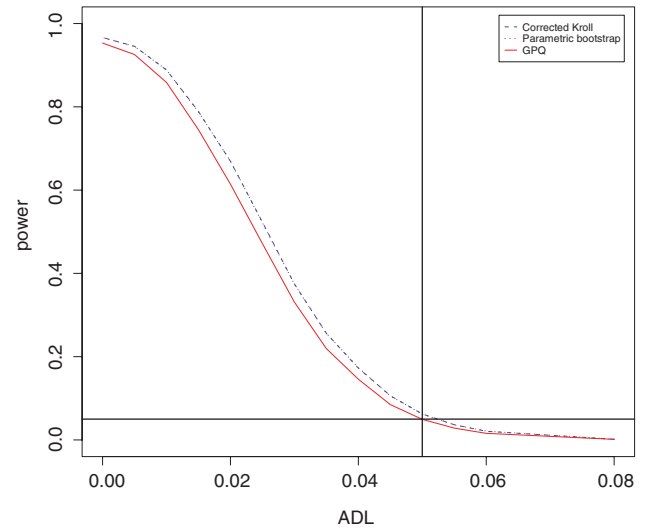

Figure 2. The empirical powers when standard deviation of normal random error is 0.2, number of solutions is 5, and number of replicates is 3. This figure is available in colour online at www.interscience.wiley.com/journal/cem

3.1. Numerical examples

The duplicate determinations at the first five concentrations given in Example 2 of CLSI guideline EP6-A [4] for assessing linearity of a new analytical procedure are used to illustrate the proposed testing procedures. The data are presented in Table III. As indicated in EP6-A [4], the criteria of $|\mu_{Pi} - \mu_{Li}|$ for claiming linearity is set as 0.2 mg/dL. For the purpose of the illustration, the allowable margin of per cent bound for ADL is set as 0.05 for the corrected Kroll's method. On the other hand, the allowable limit of SSDL for parametric bootstrap and GPQ methods is set as 0.2 which is calculated by square of 0.2 mg/dL multiplying 5 concentrations. Table IV provides the results of regression analyses for the linear, quadratic and cubic linear regression models. The results of the regression analyses presented in Table IV demonstrate that the estimates of the second-order

Table III. Measurement of calcium

Dilution	Replicate 1	Replicate 2
1	4.7	4.6
2	7.8	7.6
3	10.4	10.2
4	13.0	13.1
5	15.5	15.3

Source: The approved CLSI guideline EP6-A (2003).

regression coefficients of the quadratic model are statistically significantly different from 0 at the 5% level ($t_{0.025,\ 7} = 2.4469$) while none of them is significantly different from 0 for the cubic model. In addition, the standard error of the residuals from the estimated quadratic regression equation is 0.124 that is 39% smaller than those from the linear model. Furthermore, the coefficient of determination, R^2, is also above 0.99. As a result, the quadratic model is the best-fitted model among the three models recommended by the approved CLSI guideline EP6-A [4]. Figure 3 presents the fitted quadratic, linear regression equations, and the means at each of the five concentrations. It shows that the relationship between the dilutions of concentrations and the analytical results is nonlinear and the quadratic model is a better fit than the simple linear regression model.

Table V gives the observed predicted means from the quadratic and linear regression models at each of the five dilutions as well as their corresponding differences. Table VI presents the results of each statistical testing procedure. From these differences and the observed mean concentrations, the observed ADL yields a value of 0.0146. With respect to the hypothesis in Equation (4) and a margin of per cent bound of 5%, the critical value in Equation (5) is 0.0434 which is greater than the observed ADL of 0.0146, According to the decision rule of the corrected Kroll method, the analytical method can be concluded linear at the 5% significance level.

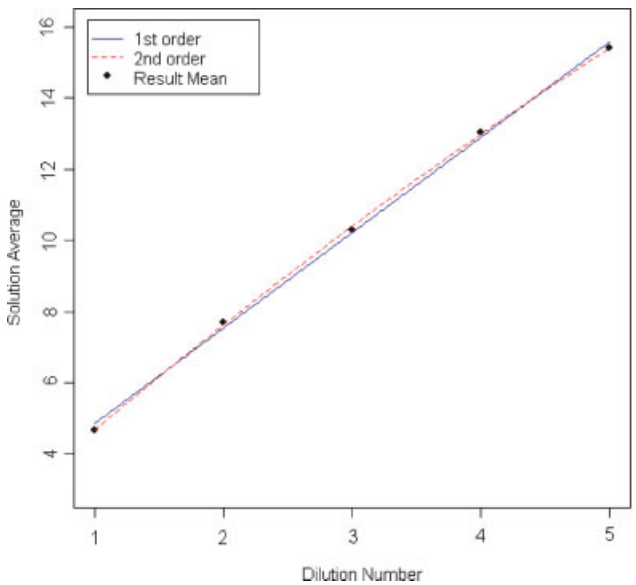

Figure 3. Regression curves for quadratic versus linear models of calcium. *Source*: The approved CLSI guideline EP6-A (2003)

Table IV. Summary of results of regression analyses

Order	Coefficient	LS Estimate	SE	*t*-test	SE $S_{y.x}$	Degrees of freedom
Linear	α'	2.16	0.15	14.3		
	β_1'	2.68	0.05	59.0	0.204	8
Quadratic	α''	1.54	0.19	8.2		
	β_1''	3.22	0.14	22.4		
	β_2''	−0.09	0.02	−3.8	0.124	7
Cubic	α'''	1.47	0.47	3.15		
	β_1'''	3.32	0.61	5.45		
	β_2'''	−0.13	0.23	−0.56		
	β_3'''	0.004	0.02	0.17	0.134	6

Source: The approved CLSI guideline EP6-A (2003).

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

Result mean	Predicted (linear)	Predicted (quadratic)	Difference	% Difference
4.65	4.85	4.67	−0.18	−3.9
7.70	7.54	7.62	0.08	1.0
10.30	10.22	10.40	0.18	1.8
13.05	12.90	12.99	0.09	0.7
15.40	15.59	15.41	−0.18	−1.2

Source: The approved CLSI guideline EP6-A (2003).

On the other hand, the 95% upper limit confidence limit for SSDL of the parametric bootstrap and GPQ methods are 0.2388 and 0.2664, respectively. As a result, both methods cannot conclude that the analytical procedure is linear at the 5% significance level. However, the 95% upper confidence limit for SSDL of the GPQ method is greater than that of parametric bootstrap method. The results presented above show the consistent results with the simulation results in Section 3.1 which the GPQ method is more conservative than other methods. However, as demonstrated by the simulation, the GPQ method is the only procedure that can adequately control the size at the nominal level.

3.2. Discussion and conclusion

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 aggregate criterion of ADL, Kroll *et al.* [1] improperly 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 of conclusion for linearity. Hsieh and Liu [2] proposed the corrected Kroll's procedures by reformulating hypothesis for proving linearity as the null hypothesis. However, the shortcoming about the unknown parameters μ and σ that need to be estimated is still required. As a result, we proposed the exact test procedures based on the aggregate criterion of SSDL. Simulation results show that the empirical size and powers of the corrected Kroll's and parametric bootstrap methods are almost the same. This phenomenon may result from the variances of the distribution for both methods are estimated by the same estimator of residual mean square obtained from the best-fitted polynomial model. On the other hand, the GPQ method not only can adequately control the type I error rate at the nominal level but also maintain the satisfactory performance of the power while the others method cannot. Therefore, we recommend the proposed statistical hypothesis in Equation (7) based on the aggregate criteria SSDL in conjunction with the testing procedure derived from the GPQ method for evaluating the linearity in assay validation. A Fortran program complied by Compaq Visual Fortran 6.6 for implement-

Table VI. Results of the linearity evaluation by three different methods

Method	Sample statistic/ Critical value or allowable bound		Conclusion
Corrected Kroll	Sample ADL	0.0146	
	Critical value	0.0434	Linear
Parametric bootstrap	Upper 95% C.L.	0.2388	
	Allowable upper bound	0.2	Quadratic
GPQ	Upper 95% C.L.	0.2664	
	Allowable upper bound	0.2	Quadratic

95% C.L.: Upper 95% Confidence limit. of SSDL.

ing linearity evaluation based on the GPQ method, the bootstrap approach and the corrected Kroll's method are available for the authors upon request.

Acknowledgements

This research is partially supported by the Taiwan National Science Council Grant: NSC 95 2118-M-002-007-MY2 to Jen-pei Liu.

REFERENCE

1. Kroll MH, Præstgaard J, Michaliszyn E, Styer PE. Evaluation of the extent of nonlinearity in reportable range studies. *Arch. Pathol. Lab. Med.* 2000; **124**: 1331–1338.
2. Hsieh E, Liu JP. On statistical evaluation of linearity in assay validation. *J. Biopharm. Stat.* 2008; **18**: 677–690.
3. ICH Expert Working Group. *International Conference on Harmonization Tripartite Guideline Q2A: Test on Validation of Analytical Procedures*, 1995.
4. Tholen DW, Kroll M, Astles JR, Caffo AL, Happe TM, Krouwer J, Lasky F. *EP6-A: Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline,* Clinical Laboratory Standard Institute, Wayne, PA, U.S.A., 2003.
5. Tsui KW, Weerahandi S. Generalized *p*-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Am. Stat. Assoc.* 1989; **84**: 602–607.
6. Weerahandi S. Generalized confidence intervals. *J. Am. Stat. Assoc.* 1989; **88**: 899–905.
7. McNally RJ, Iyer HK, Mathew T. Tests for individual and population bioequivalence based on generalized *p*-values. *Stat. Med.* 2003; **22**: 31–53.
8. Liao CT, Iyer HK. A Tolerance interval for the normal distribution with several variance components. *Stat. Sin.* 2004; **14**: 217–229.
9. Liao CT, Lin TY, Iyer HK. One- and two-sided tolerance intervals for general balanced mixed models and unbalanced one-way random models. *Technometrics* 2005; **47**: 323–335.
10. Li CR, Liao CT, Liu JP. On the exact interval estimation for the difference in paired areas under the roc curves. *Stat. Med.* 2008; **27**: 224–242.
11. Li CR, Liao CT, Liu JP. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Stat. Med.* 2008; **27**: 1762–1776.
12. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, New York, USA 1993.
13. Searle SR. *Linear Models*. Wiley: New York, New York, U.S.A., 1971.
14. Hanning J, Iyer HK, Patterson P. Fiducial generalized confidence intervals. *J. Am. Stat. Assoc.* 2006; **101**: 254–269.