Evaluation of the Extent of Nonlinearity in Reportable Range Studies

Martin H. Kroll, MD; Jens Præstgaard, PhD; Ellen Michaliszyn, BS; Patricia E. Styer, PhD

• ***Objectives.*—To extend the polynomial method for evaluating linearity in 2 ways. First, we developed a screen to ascertain whether the data were precise enough to permit a reliable evaluation of linearity and therefore eliminate findings of linearity due to low statistical power. Second, we assessed whether the degree of nonlinearity detected by the polynomial method was clinically relevant using a statistically rigorous method.**

***Methods.*—Because we assessed linearity relative to a clinically determined level of importance instead of the default value of zero, we used sampling theory based on the noncentral χ^2 distribution. Using statistical power calculations, we incorporated a screen for imprecision that guarantees that the probability of correctly identifying nonlinear methods is at least 80%.**

***Results.*—With the described methods, we achieved a sensitivity of at least 80% and a specificity of at least 95%. When the data were too imprecise to achieve a sensitivity of 80%, no determination of linearity was made. This procedure mimics the practice in manual inspection of flagging data that appear imprecise by visual inspection and halting the evaluation.**

***Conclusions.*—Formal statistical tests for precision and amount of nonlinearity are advantageous because they allow us to quantify and limit classification errors. By formalizing these various aspects of linearity assessment, we maintain some of the complex features of manual methods while making the linearity assessment feasible to apply to a high volume of assessments and removing the between-analyst variability.**

(*Arch Pathol Lab Med.* 2000;124:1331–1338)

The College of American Pathologists (CAP) provides Linearity and Calibration Verification Surveys to assess the linearity of participating laboratories. In these Surveys, the laboratories receive a number of serum samples spiked in increasing concentrations with the analyte to be investigated. The laboratories then measure the analyte in each sample in duplicate and submit the measurement to the CAP for data processing. The data-processing procedure uses linear regression analysis and significance tests to determine whether the participants' data are best modeled by a straight line or a nonlinear curve in the form of a second- or third-order polynomial.

The basis of the polynomial method is to use linear regression techniques to find the polynomial that best fits the data.[1,2] The methods start out by fitting a polynomial of sufficiently high order, for example a cubic polynomial. The *P* value for the highest order coefficient is investigated by a *t* test; if this coefficient is not significant, then a polynomial of 1 degree less is fitted. The process is then repeated until no further reductions in order can be made; the resulting polynomial is considered to be the best fit to the data. If the best-fit polynomial is a straight line, then the data and the method producing them are called linear; if not, the data and method are called nonlinear.

In previous work, a distinction was made between clinically relevant and clinically unimportant findings of nonlinearity.[1] One innovation in our current methodology is a refinement of the assessment of clinically relevant nonlinearity using more rigorous statistical sampling theory. We propose a simple method to quantify the amount of nonlinearity and compare it to an ''allowable difference,'' which is determined relative to the goal for total error for the analyte in question. In a usual statistical test, observed deviations from linearity are compared to a hypothesized difference of zero. By assessing the amount of nonlinearity to a lower bound other than zero, we introduce some nonstandard distribution theory. The advantage of doing this in a statistically rigorous manner is that we are then able to evaluate and limit misclassification errors given a known set of assumptions.

The polynomial method was a major breakthrough in linearity evaluation in that it enabled the CAP to automatically perform a high-volume and consistent evaluation for linearity of participant data. After applying the polynomial method, it became clear that adding a constraint on the amount of allowable imprecision would improve the assessment of linearity. As described, the polynomial method uses significance tests to determine if a second- or third-degree polynomial fits the data better than a straight line. If the data have poor precision, then the statistical power of these significance tests will be low, resulting in a default evaluation of ''linear'' for the imprecise data set. The second innovation in the present article is to add a screen for imprecise data, and hence guarantee

Accepted for publication March 31, 2000.

From the Department of Pathology, Dallas Veterans Affairs Medical Center, Dallas, Tex (Dr Kroll); Abbott Laboratories, Abbott Park, Ill (Dr Præstgaard); and the College of American Pathologists, Northfield, Ill (Ms Michaliszyn and Dr Styer).

Reprints: Martin H. Kroll, MD, Department of Pathology, Dallas Veterans Affairs Medical Center, 4500 Lancaster Rd, 113, Dallas, TX 75216.

a certain level of statistical power. Imprecise data do not contain information to prove or disprove linearity, so the prudent outcome is not to evaluate such data. This is a problem for any method of linearity evaluation, especially visual or discrete methodologies.

Goals of the CAP surveys have been to minimize the number of errant evaluations and to make the method as objective as possible. To be objective, a method must minimize the necessity for visual inspection of the data points in a graphic format. The goal should be that different evaluators, applying the same method, should always reach the same conclusion.

In this article, we present an extension of the polynomial method, allowing us to give a better evaluation of imprecise data sets and utilize a more rigorous assessment of clinical relevance for statistically nonlinear data sets. We still begin with a regression analysis to determine if a higher-order polynomial fits data better than a straight line. We then evaluate 2 additional constraints. First, the data must be precise enough to guarantee reasonable statistical power. If the data are too imprecise, then no final determination of linearity is given. Second, if the data appear to be nonlinear after the evaluation of the best-fit polynomial, then the amount of nonlinearity is assessed using criteria reflecting clinical relevance. Evaluations of both precision and clinical relevance are made using formal statistical tests. This permits the quantification of classification errors. In the first case, by screening for data that are too imprecise, we guarantee that at least 80% of the truly nonlinear data sets will be identified correctly. In the second case, we limit the misclassification of truly linear data sets to 5% using the standard of clinical relevance instead of the usual evaluation of statistical significance relative to zero.

METHODS

For carrying out the linearity protocol, measure S solutions R times (S represents the number of solutions and R the number of repetitions) and find the best-fitting polynomial using the method of Kroll and Emancipator.[1] We denote this polynomial as $p(x)$, where x is in the set of S solutions. Typically, $p(x)$ is a linear, quadratic, or cubic polynomial. The estimated precision from the best-fit polynomial, defined to be the mean square error and denoted by σ, also plays an important role in the linearity evaluation. Methods with high imprecision show large differences between the observed values and the estimated best-fit polynomials.

Our method stipulates that when the polynomial $p(x)$ is not linear, we must determine whether the difference from the best-fit straight line is large enough to be clinically significant. We call our measure of nonlinearity the *average deviation from linearity* (*ADL*), and define it as

$$ADL = \frac{\sqrt{\sum_{x \in X} [p(x) - (a + bx)]^2 / S}}{\bar{c}},$$

where $p(x)$ is the best-fit polynomial, $a + bx$ is the best-fit straight line (ie, the simple regression line), $\bar{c}$ is the mean concentration for all solutions of the assay, and the summation is taken over each of the S solution levels.[1,2] The *ADL* is the square root of the average squared distances between the fitted point on the best-fit polynomial curve and the simple regression line for each solution level, standardized by dividing by the mean concentration $\bar{c}$. Figure 1 presents a graphic example of how the *ADL* is calculated for a typical linearity evaluation. We selected this particular measure of nonlinearity because it is convenient to put all evaluations on the same coefficient of variation scale, and it is straightforward to derive the corresponding sampling distribution based on the standard theory of linear models.[3] Appendix 1 contains additional details of the sampling distribution.

In assessing the clinical relevance of any observed nonlinearity, we must evaluate the size of the *ADL*. As previously stated, we will not compare the *ADL* with a hypothesized difference of zero. Rather, if the average deviation from linearity is less than some percent bound, we conclude that the observed nonlinearity is not clinically significant. For many analytes, a percent bound of 5% would be reasonable, but this bound can be modified depending on the expected or clinically required accuracy of the measurements, or it can be based on biological variation.[4] In this study, we use a percent bound of 5%. We denote the cutoff for clinical relevance as *PctBnd*.

The average deviation from linearity is a statistic calculated from a random sample of measurements, so we must consider the underlying sampling distribution of the *ADL* when evaluating its size relative to any specified percent bound. We are in fact carrying out a formal statistical test of the *ADL* when we evaluate the clinical relevance of any observed nonlinear result. Similarly, we will build on this same statistical theory to screen for methods with high imprecision.

In analogy with diagnostic tests, true system linearity or nonlinearity corresponds to the absence or presence of a disease. Here, system linearity is defined as the ''true'' value of the average deviation from linearity being less than the *PctBnd*. The ''true'' value is the value we would get for the average deviation from linearity in the hypothetical situation that the imprecision is zero (ie, that there is no measurement error so the observed points fall precisely on a curve determined by the best-fit polynomial). Our ''diagnostic test'' for linearity classifies the system as linear if the observed average deviation from linearity is less than the appropriate critical value from the sampling distribution of the *ADL*. Table 1 shows the situations that can occur. The sensitivity of the linearity evaluation is the percentage of true positives, that is, the percentage of evaluations that show nonlinear results when the system is in fact nonlinear. The specificity of the evaluation is the percentage of true negatives, or the percentage of evaluations that would result in a linear classification when the system is truly linear.

It is not possible to have 100% sensitivity and specificity for any classification system. In fact, requiring weaker evidence for nonlinearity to maximize the sensitivity will lead to missing more nonlinear classifications, and hence will lower the specificity.[5] We chose a formal statistical evaluation of linearity precisely so we can characterize the sensitivity and specificity of our linearity evaluations, or conversely, our misclassification errors. We followed general conventions in setting the lower bound for the specificity at 95% (corresponding to a type I error of 5%) and the sensitivity at 80% (corresponding to a type II error of 20%). Appendix 1 outlines some of the distribution theory used to calculate the necessary constants for the evaluation of linearity. The results are presented without detailed explanations in Tables 2, 3, and 4.

Screen for Imprecision

Once the best-fit polynomial is determined, we screen for imprecision. First, the estimated precision from the best-fit polynomial must be calculated. The estimated precision, denoted σ, is the standard deviation around the best-fit polynomial, calculated as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} [y_i - p(x_i)]^2}{n - d - 1}}, \quad (1)$$

where y_i represents the observed results, $p(x_i)$ represents the fitted values from the best-fit polynomial, n is the total number of observations $S \cdot R$, and d is the degree of the best-fit polynomial. The imprecision screen is based on the magnitude of $\sigma/\bar{c}$, the estimated precision from the best-fit polynomial divided by the mean concentration for all of the assay solutions. To evaluate the

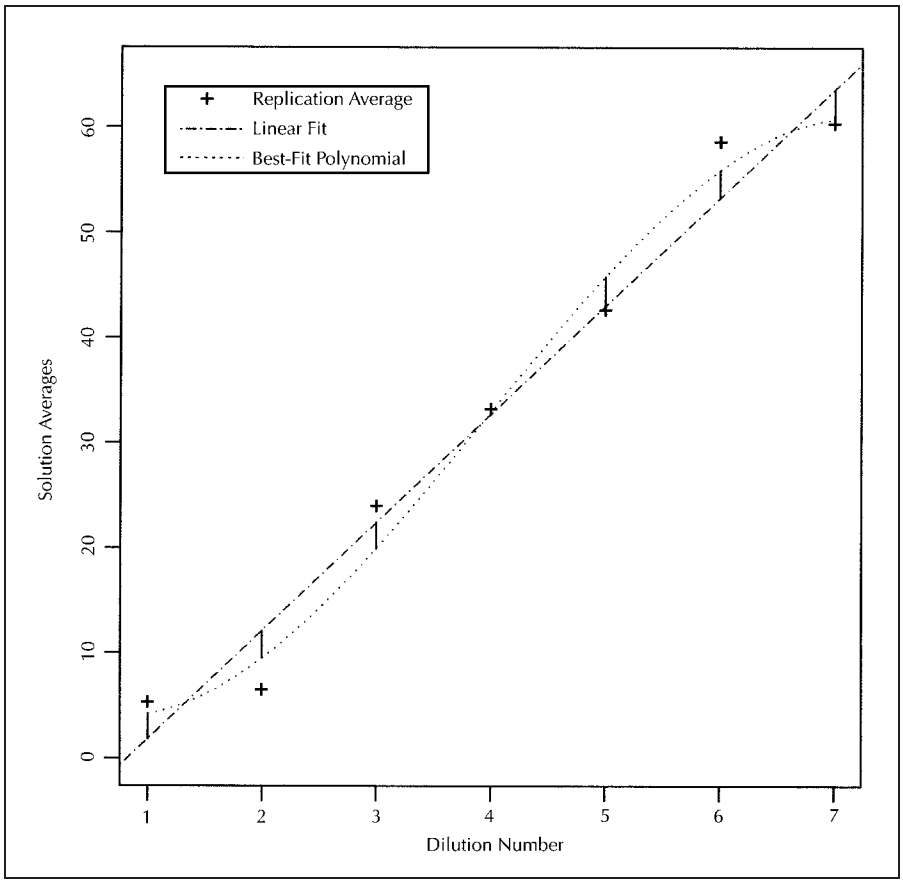

Figure 1. *Derivation of average deviation from linearity (*ADL*).*

Table 1. Possible Outcomes of the Linearity Experiment

Conclude From Observations	True Linearity Category	
	System is Nonlinear	System is Linear
Classification is nonlinear	True positive Correct conclusion $(1 - \beta)$ = Sensitivity	False positive Type I error α
Classification is linear	False negative Type II error β	True negative Correct conclusion $(1 - \alpha)$ = Specificity

Table 2. Constant for Calculating the Imprecision Bound to Ensure Specificity of 95% and Sensitivity of 80%

Degree of Best-Fit Polynomial	Constant (C) for Bound on Precision
Linear or quadratic	6.3
Cubic	6.5

imprecision, the ratio $\sigma/\bar{c}$ is compared to a quantity based on the limit set for clinical relevance, the total number of measurements, and a constant that depends on the degree of the best-fit polynomial and the desired level of specificity and sensitivity. The condition necessary to guarantee sensitivity of 80% with a corresponding specificity of 95% is

$$\frac{\sigma}{\bar{c}} < PctBnd\sqrt{\frac{S \cdot R}{C}}, \quad (2)$$

where $S{\cdot}R$ is the number of solutions times the number of replications, and C is a constant given in Table 2. As noted, derivation of C depends on the underlying distribution theory and is outlined in Appendix 1.

Evaluate Nonlinearity Relative to Clinical Relevance

To determine if the assay is clinically linear given a quadratic or cubic best-fit polynomial, we compare the observed *ADL* to a critical value derived from the appropriate sampling distribution for the *ADL*. The sampling distribution for the *ADL* (assuming that the true deviation from linearity is small but not zero) is the noncentral χ^2 distribution.

A statistical test is carried out by comparing an observed test statistic to a critical value of the sampling distribution of the test statistic. The critical value depends on the stated specificity of the test, usually 95%. For our test, the critical value is the 95th quantile of the noncentral χ^2 distribution, with noncentrality parameter dependent on *PctBnd,* the stated standard for clinical relevance. The 95th quantile of the noncentral χ^2 distribution is the value such that 95% of the observations are less than that value and 5% are greater. With denoting the 95th quantile of the appropriate noncentral χ^2 distribution, the critical value for the observed *ADL* is

$$\frac{\sigma}{\bar{c}}\sqrt{\frac{q_{0.95}}{S \cdot R}}, \quad (3)$$

where $S{\cdot}R$ is the total number of observations and $\sigma/\bar{c}$ is the imprecision estimated from the best-fit polynomial divided by the overall mean of the solutions. If the observed *ADL* is less than the value given by equation (3), then the results are classified as linear. This means that although there was a statistically significant nonlinearity detected by the Kroll and Emancipator polynomial method, the degree of nonlinearity was judged to be unimportant clinically. Appendixes 1 and 2 provide more details for implementing this formula directly.

Table 3. Imprecision Screen and Critical Values for *ADL* With *PctBnd* = 5% and Degree of Polynomial = 1 or 2*						
$\sigma/\bar{c}$, %	*S·R* = 10	*S·R* = 12	*S·R* = 14	*S·R* = 16	*S·R* = 18	*S·R* = 20
1	5.5	5.5	5.4	5.4	5.4	5.4
2	6.1	6.0	5.9	5.8	5.8	5.7
3	6.6	6.4	6.3	6.3	6.2	6.1
4	7.1	6.9	6.8	6.7	6.6	6.5
5	6.6	7.4	7.2	7.1	7.0	6.9
6	8.2	7.9	7.7	7.5	7.4	7.2
7	8.7 (P)	8.4 (P)	8.1	7.9	7.8	7.6
8	P	P	8.6 (P)	8.3 (P)	8.1	8.0
9	P	P	P	P	8.5 (P)	8.3 (P)
>9	P	P	P	P	P	P

* $\sigma/\bar{c}$ is the error coefficient of variation expressed as a percent. *S·R* is the number of solution levels times the number of replications. P indicates that the results corresponding to these cells are too imprecise to evaluate. Use linear interpolation to estimate the critical values that fall between the tabled values of $\sigma/\bar{c}$.

Table 4. Imprecision Screen and Critical Values for *ADL* With *PctBnd* = 5% and Degree of Polynomial = 3*						
$\sigma/\bar{c}$, %	*S·R* = 10	*S·R* = 12	*S·R* = 14	*S·R* = 16	*S·R* = 18	*S·R* = 20
1	5.5	5.5	5.4	5.4	5.4	5.4
2	6.1	6.0	5.9	5.9	5.8	5.8
3	6.7	6.5	6.4	6.3	6.2	6.2
4	7.2	7.0	6.9	6.8	6.7	6.6
5	7.8	7.6	7.4	7.2	7.1	7.0
6	8.4	8.1	7.9	7.7	7.5	7.4
7	9.0 (P)	8.7 (P)	8.4	8.2	8.0	7.8
8	P	P	8.9 (P)	8.6 (P)	8.4	8.2
9	P	P	P	P	8.9 (P)	8.7 (P)
>9	P	P	P	P	P	P

* $\sigma/\bar{c}$ is the error coefficient of variation expressed as a percent. *S·R* is the number of solution levels times the number of replications. P indicates that the results corresponding to these cells are too imprecise to evaluate. Use linear interpolation to estimate the critical values that fall between the tabled values of $\sigma/\bar{c}$.

Tables 3 and 4 show the critical values based on a specificity of 95% and include a way to screen for imprecision that bypasses equation (2). The tables use *PctBnd* = 5% for judging clinical importance. The horizontal entry in each table is indexed by the number of observations in the experiment (*S·R*) and the vertical entry by the ratio of the precision to the mean concentration, $\sigma/\bar{c}$, expressed as a percentage. In both tables, we can see that the critical value increases as $\sigma/\bar{c}$ increases. The critical value also decreases as the number of observations increases. Cells marked with a ''P'' are cases in which the imprecision is too great to complete the linearity evaluation with 80% sensitivity. At these values, the data are considered too imprecise to evaluate.

Table 3 can also be used to determine if the imprecision is too great to confirm the finding of linearity when the best-fit polynomial is a straight line. Here, the observed *ADL* is zero, so the critical values are not useful. Rather, when the imprecision as measured by $\sigma/\bar{c}$ is large, the ''P'' notation indicates that the data are too imprecise to permit the finding of linearity. Checking to see if the corresponding cell of the table contains a ''P'' is equivalent to evaluating the inequality in equation (2).

In summary, Tables 3 and 4 can be used in conjunction with Kroll and Emancipator's polynomial method as follows:

- Compute the mean solution $\bar{c}$.
- Perform polynomial regressions to determine the order of polynomial that best fits the data (linear, quadratic, or cubic).
- If the best fitting polynomial is quadratic or cubic, compute the average deviation from linearity. Otherwise, set the average deviation from linearity to zero.
- From the regression output, estimate the precision σ and compute $\sigma/\bar{c}$. The estimate for σ is the root mean squared error of the best fitting polynomial, as defined in equation (1). If using SAS PROC REG, these values are given in the default printed output, labeled ''Root MSE,'' ''Dependent Mean,'' and ''Coeff Var.''
- Select the appropriate table and locate the cell corresponding to the approximate $\sigma/\bar{c}$ and the given number of observations. If the table contains a ''P,'' then the data are too imprecise to evaluate.
- If the best-fit polynomial is linear and the corresponding table entry is not a ''P,'' then the system is linear.
- If the best-fit polynomial is nonlinear and the tabled entry is not a ''P,'' compare the average deviation from linearity with the tabled critical value. Use linear interpolation to estimate the critical values that fall between the tabled values. If the average deviation from linearity exceeds the critical value, then the system is nonlinear. Otherwise, the system is linear.

Appendixes 1 and 2 provide more details on the derivation of these tables and how to implement them for linearity evaluations.

ILLUSTRATION OF THE EVALUATION METHODOLOGY

In this section, we show how the proposed evaluation works on measurements from the CAP Linearity Survey LN2-1999A. The measurements of 7 concentrations shown (each measured in duplicate) are all for the analyte lactate dehydrogenase. In all cases, we use a nominal cutoff value *PctBnd* = 5%, and we take $\bar{c}$ to be the average of the submitted measurements. Table 5 illustrates the decision tree used for each example.

Example 1: Linear 1

Figure 2, A, shows a set of measurements that appear to fall on a straight line. When we fit second- and third-order polynomials by linear regression, none of the higher-order coefficients are significant. We conclude that the system is best described by a line, and that the average deviation from linearity is zero. To check that the imprecision is acceptable, we compute the ratio of precision to

Table 5. Outline of Decision Rules for Linearity Evaluation

Case 1: The best-fit polynomial is of degree 1 (linear)

Check data for imprecision: Is $\sigma/\bar{c}$ greater than the limit on precision?
- If yes, then data are too imprecise to evaluate.
- If no, then data are linear.

Not necessary to check for clinical relevance.

Case 2: The best-fit polynomial is nonlinear

Check data for imprecision: Is $\sigma/\bar{c}$ greater than the limit on precision?
- If yes, then data are too imprecise to evaluate.
- If no, then check for clinical relevance.

Check for clinical relevance: Is *ADL* greater than the critical value?
- If yes, then data are nonlinear.
- If no, then data are linear.

mean concentration: σ = 44.78 IU/L, $\bar{c}$ = 2302.2 IU/L, so $\sigma/\bar{c}$ = 1.9%. Since the polynomial is a line of degree d = 1, we look in Table 3 under 14 observations and $\sigma/\bar{c}$ between 1% and 2%. The value in the table is between 5.4% and 5.9%. Since these values do not have the ''P'' entry, we conclude that the data can be evaluated for linearity, and we assign the ''Linear 1'' evaluation.

Example 2: Linear 2

Figure 2, B, shows a data set in which a second-order polynomial is determined to be the best-fit polynomial. The mean concentration is $\bar{c}$ = 1262.5 IU/L, and since we use a curve rather than a straight line to describe the data, we compute the average deviation from linearity to be 1.2%. The ratio of precision to mean concentration is $\sigma/\bar{c}$ = 1.3%. Since the polynomial is of order 2, we look in Table 3. The entry corresponding to $\sigma/\bar{c}$ = 1.3% and 14 observations is between 5.4% and 5.9%. In this case, the tabled critical value is larger than the observed average deviation from linearity. Since the table contains no ''P'' for this entry, we give the evaluation a ''Linear 2'' designation, concluding that the observed nonlinearity in the data is not clinically relevant.

Example 3: Nonlinear

For the data in Figure 2, C, a third-order polynomial is selected as the best-fit polynomial, and we calculate an average deviation from linearity of 8.6%. The precision at the mean concentration is $\sigma/\bar{c}$ = 5.5%. From Table 4, the interpolated value between $\sigma/\bar{c}$ = 5% and $\sigma/\bar{c}$ = 6% is 7.7 for 14 observations. Since the observed *ADL* is greater than the interpolated value from the table, we conclude that the results are nonlinear. See Appendix 2 for a detailed illustration of the calculations for this example.

Example 4: Too Imprecise

In Figure 2, D, a straight line is selected as the best-fit polynomial. Accordingly, the average deviation from linearity is zero. The precision is estimated to be 36.2%. By looking in Table 3 under 14 observations and $\sigma/\bar{c} > 10\%$,

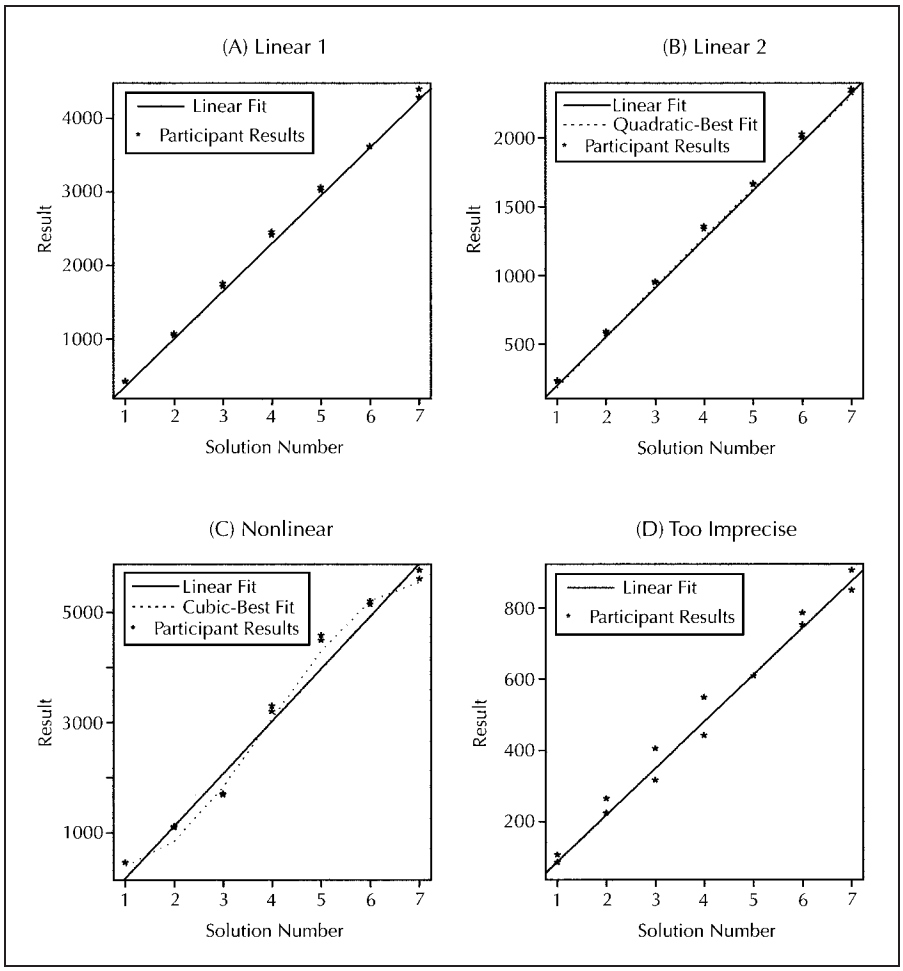

Figure 2. *Examples of possible results of linearity evaluation.*

we see that we are in a cell marked ''P.'' We conclude that the data are too imprecise to evaluate for linearity.

COMMENT

Knowledge of the specificity and sensitivity of the method for detecting linearity has important implications for the use of any method for evaluating linearity. Any method for evaluating linearity is a test of whether the system is linear. As with any test, the results fall into 3 categories: linear, nonlinear, and too difficult to decide. Likewise, all tests for linearity have implicit misclassification errors. Unlike many other methods for testing linearity, our method permits the estimation of misclassification errors. The sensitivity and specificity have strict lower bounds that do not depend on the precision of the data. Methods that are not based on quantitative approaches or that are subjective in nature do not calculate specificity and sensitivity. Therefore, such methods cannot stipulate a successful interpretation of linearity test results.

The specified minimum specificity of 95% and sensitivity of 80% describe the performance characteristics in a single evaluation of linearity. Because neither the specificity nor sensitivity are 100%, evaluations for linearity will include false positives and false negatives. The probability that the described method would declare a linear method as nonlinear twice in a row is 4 out of 1000, and 3 times in a row, less than 2 out of 10 000. The sensitivity limit implies one must test for nonlinearity at least twice to achieve a 95% sensitivity, and 3 times to achieve a 99% sensitivity. Therefore, one should check an analytical method for nonlinearity more than once to assure one's methods are linear. Furthermore, testing 3 times at intervals of every 6 months usually will encompass the usable lifetime of a calibrator lot and control material lot combination life span. Therefore, following an initial 3 evaluations of linearity, continued testing, on the order of every 6 months, is recommended.

APPENDIX 1

Theoretical Development of Linearity Evaluation

Testing in linear models is typically done with an F test. The F distribution is the ratio of 2 χ^2 variables and their corresponding degrees of freedom. Our definition of the *ADL* is a simple function of the χ^2 variable that forms the numerator of an F statistic for comparing 2 models in the general linear model theory.[3] Specifically, we have fit 2 models, the polynomial model and the straight-line model. The polynomial model is fit in a $(d + 1)$-dimensional linear subspace of R^n, and the straight-line model is fit in a linear subspace of that subspace. Standard notation for the χ^2 variable of interest is $[\Sigma_{ij}(y_{ij} - \hat{\mu}_{oi})^2 - (y_{ij} - \hat{\mu}_i)^2 \sim \sigma^2 \cdot \chi^2(d - 1,\theta)$, where y_{ij} represents the observed results, $\hat{\mu}_{oi}$ is the corresponding fitted value for the ith solution from the straight-line model, and $\hat{\mu}_i$ is the corresponding fitted value from the polynomial model. This approach is also known as the decomposition of the variability of the observed y_{ij}'s, and we are focusing the component that would be called the sum of squares from the polynomial terms in the model. An equivalent expression for the sum of squares from the polynomial terms, which is more convenient for our application, is

$$\left[\sum_i (\hat{\mu}_i - \hat{\mu}_{oi})^2\right] \sim \frac{\sigma^2}{R} \cdot \chi^2(d - 1, \theta), \qquad (1.1)$$

where R is the number of replicates within each solution level. The noncentrality parameter θ is a function of the expected sum of squared deviations between the 2 sets of fitted values. In the case of replicated measurements with R replicates per solution level, the textbook version of the noncentrality parameter is $\theta = \Sigma_i(\mu_i - \mu_{oi})^2/(\sigma^2/R)$, where i indexes the set of solution values.

As noted in the text, we want to specify the degree of nonlinearity as a percent bound so that it is easier to specify and interpret. To accomplish this, we call the allowable deviation from linearity under the null hypothesis adl_{true}, and express it as the sum of squared deviations divided by the number of solution levels, and standardized by the overall mean. Hence,

$$adl_{true} = \left(\frac{1}{\bar{c}}\right) \cdot \sqrt{\sum_i (\mu_i - \hat{\mu}_i)^2/S}\,.$$

Note that this is the expected value of the *ADL* as stated in the text.

Let *SSD* denote the left-hand side of equation (1.1). Then, the *ADL* is simply

$$\left(\frac{1}{\bar{c}}\right) \cdot \sqrt{SSD/S}\,,$$

where $\bar{c}$ is the mean of the observed results and S is the number of unique solution values. Also, note that the distribution in equation (1.1) can be re-expressed as

$$SSD \sim \frac{\sigma^2}{R} \cdot \chi^2\left[d - 1, \frac{(adl_{true} \cdot \bar{c})^2 \cdot S}{\sigma^2/R}\right].$$

Letting *PctBnd* equal the specified value of adl_{true} under the null hypothesis, this can be rewritten as

$$SSD \sim \frac{\sigma^2}{R}\chi^2\left[d - 1, \frac{(PctBnd)^2 \cdot S \cdot R}{\left(\frac{\sigma}{\bar{c}}\right)^2}\right], \qquad (1.2)$$

where $d - 1$ is the degrees of freedom and the second term in the large brackets is the noncentrality parameter as specified for our application. Since the *ADL* is a simple function of the *SSD,* the critical value for determining if an observed nonlinear result meets the criteria of clinical importance is also a simple function of the 95th quantile of the sampling distribution for the *SSD.* Let $q_{0.95}$ denote the 95th quantile of this noncentral χ^2 distribution, so the 95th quantile for the *SSD* would be $(\sigma^2/R) \cdot q_{0.95}$. The 95th quantile of the sampling distribution for the *ADL* can be expressed as

$$\text{Critical Value} = \frac{\sigma}{\bar{c}}\sqrt{\frac{q_{0.95}}{S \cdot R}}, \qquad (1.3)$$

where $S \cdot R$ is the number of observations, $\sigma/\bar{c}$ is the standardized precision, and $q_{0.95}$ is the 95th quantile of the appropriate noncentral χ^2 distribution. If the underlying assumptions of the statistical test are met, we have met the condition that the specificity is at least 95%.

While we have provided tables for the critical values for a number of combinations of *S, R,* and $\sigma/\bar{c}$, all of the tabled values use 5% for the *PctBnd.* To calculate critical values for other combinations of parameters, we suggest using a simple approximation for the 95th quantile of the corresponding noncentral χ^2 distribution.[6] The 95th quantile of a noncentral χ^2 random variable can be ap-

proximated by the 95th quantile of a usual χ^2 random variable with the following multiplier and degrees of freedom:

$$q_{0.95} = \left(\frac{df + 2 \cdot ncparm}{df + ncparm}\right) \cdot \chi^2_{0.95}\left\{\left[\frac{df + ncparm)^2}{df + 2 \cdot ncparm}\right]\right\},$$

where *ncparm* is the noncentrality parameter

$$ncparm = \frac{PctBnd^2 \cdot S \cdot R}{(\sigma/\bar{c})^2},$$

and $df = d - 1$, where d is the degree of the best-fit polynomial.

Maintaining a Minimum Sensitivity

The sensitivity depends on the degree of nonlinearity that we are trying to detect. We set the amount of nonlinearity that we must be able to detect with 80% sensitivity to be twice the value for clinical relevance. That is, our criterion is that a system whose true average deviation from linearity is $2 \cdot PctBnd$ must be correctly classified with at least 80% sensitivity. Under the alternative hypothesis that the true *ADL* is $2 \cdot PctBnd$, the sampling distribution of the observed SSD has the following scaled noncentral χ^2 distribution:

$$SSD \sim \frac{\sigma^2}{R}\chi^2\left[d - 1, 4\frac{(PctBnd)^2 S \cdot R}{\left(\frac{\sigma}{\bar{c}}\right)^2}\right].$$

Note that this is the same distribution as in equation (1.2), except the noncentrality parameter is now multiplied by 4. By the same logic used above, the sampling distribution of the *ADL* is simply a rescaled noncentral χ^2 distribution, so it is straightforward to calculate the sensitivity. Letting K denote the critical value from the sampling distribution under the null hypothesis that $adl_{true} = PctBnd$, so K is equal to the critical value described in equation (1.3), the sensitivity is the area above K from the corresponding sampling distribution under the alternative hypothesis that $adl_{true} = 2 \cdot PctBnd$.

As stated in the text, given constant values for *PctBnd* and the number of observations in the linearity experiment, the condition

$$\frac{\sigma}{\bar{c}} < PctBnd\sqrt{\frac{SR}{C}} \qquad (1.4)$$

ensures that the linearity evaluation will have sensitivity of at least 80%. The values for C were derived by calculating the appropriate integrals from the noncentral χ^2 distributions under the null and alternative hypotheses for a grid of values of the noncentrality parameter for the null distribution. The constant C is in fact a lower bound on the noncentrality parameter for the null distribution.

If the best-fit polynomial is linear, we still evaluate the sensitivity based on the sampling distribution of the average deviation from linearity. Here, the observed *ADL* is zero, so we do not carry out the evaluation of clinical relevance. We use the sampling distribution for the quadratic best-fit polynomial to screen for imprecision. We use the quadratic case because it provides the most stringent screen, and thus guarantees the greatest sensitivity.

APPENDIX 2

Guide for Independent Linearity Evaluations

This sample calculation uses the data in example 3, where the result is a clinically important nonlinear evaluation. Note that the calculations are carried out on the percent scale.

Sample data:

Solution Number (x): 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7

Result (y): 352, 348, 1009, 991, 1603, 1584, 3100, 3200, 4482, 4390, 5101, 5046, 5669, 5516

Mean of y ($\bar{c}$): 3027.9

Using any standard regression package, find best-fit polynomial.

Degree of best-fit polynomial (1, 2, or 3) $d = 3$

Number of solutions and replications $S = 7$; $R = 2$

If $d = 2$ or 3,

1. Obtain the fitted values at each solution level for the best-fit polynomial $p(x) = 165.4, 1119.6, 2073.7, 3027.9, 3982.1, 4936.3, 5890.5$.
2. Obtain the fitted values at each solution level for the linear regression line $(a + bx) = 386.2, 840.6, 1829.7, 3074.4, 4295.9, 5215.2, 5553.5$.
3. Calculate the average deviation from linearity and convert to percent

$$ADL \cdot 100\% = \frac{\sqrt{\sum_{x \in X}[p(x) - (a + bx)]^2/S}}{\bar{c}} \cdot 100\%$$

$$ADL = 8.6\%$$

Estimate of regression standard error from the best-fit polynomial $\sigma = 167.8$

Estimate of solution mean $\bar{c} = 3027.9$

Estimate of precision on percent scale $\sigma/\bar{c} \cdot 100 = 5.5\%$

Criteria for clinical relevance $PctBnd = 5\%$

Bound C from Table 2 (optional) C = 6.5

Limit on precision

Is entry in Table 3 or 4 a ''P''? Yes or No No

Critical value

Entry from Table 3 or 4 (interpolated)

Critical Value = 7.65%

Optional Method by Direct Calculation

Limit on precision

$$\text{Is } \frac{\sigma}{\bar{c}} \cdot 100\% > PctBnd \cdot \sqrt{\frac{S \cdot R}{C}}?$$

$$\text{Is } 5.5\% > 5\% \cdot \sqrt{\frac{7 \cdot 2}{6.5}}? \quad \text{Yes or No No}$$

Critical Value

Step 1: Get $q_{0.95}$ for noncentral χ^2 using approximation from regular χ^2 with an estimated multiplier and degrees of freedom:

$$df = d - 1 = 2$$

$$ncparm = \frac{PctBnd^2 \cdot S \cdot R}{(\sigma / \bar{c})^2} = 11.57$$

$$q_{0.95} = \left(\frac{df + 2 \cdot ncparm}{df + ncparm}\right) \cdot \chi^2_{0.95}\left\{\left[\frac{(df + ncparm)^2}{df + 2 \cdot ncparm}\right]\right\}$$

$$= \left(\frac{2 + 2 \cdot 11.6}{2 + 11.6}\right) \cdot \chi^2_{0.95}\left\{\left[\frac{(2 + 11.6)^2}{2 + 2 \cdot 11.6}\right]\right\}$$

$$= 1.85 \cdot \chi^2_{0.95}(7.3)$$

$$\chi^2_{0.95}(7.3) = 14.5, \quad \text{so} \quad q_{0.95} = 1.85 \cdot 14.5 = 26.9$$

Step 2:

$$\text{Critical Value} = \frac{\sigma}{\bar{c}}\sqrt{\frac{q_{0.95}}{S \cdot R}} = 7.62\%$$

Note. Find $\chi^2_{0.95}$ (7.3) using software that permits specification of noninteger degrees of freedom or round off and use a published table to find the 95th quantile of the χ^2 distribution with 7 df; in EXCEL, use gammainv(.95,7.3/2,2); in SAS use 2·gaminv(.95,7.3/2). Other programs may utilize different paramaterizations of the gamma distribution or estimate noninteger degrees of freedom directly for the χ^2 distribution.

References

1. Kroll MH, Emancipator K. A theoretical evaluation of linearity. *Clin Chem.* 1993;39:405–413.
2. Emancipator K, Kroll MH. A quantitative measure of nonlinearity. *Clin Chem.* 1993;39:766–772.
3. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco, Calif: Holden-Day; 1977.
4. Fraser CG. The application of theoretical goals based on biological variation data in proficiency testing. *Arch Pathol Lab Med.* 1988;112:404–415.
5. Glantz SA. *Primer of Biostatistics.* New York, NY: McGraw-Hill Inc; 1987.
6. Winer BJ, Brown DR, Michels KM. *Statistical Principles in Experimental Design.* New York, NY: McGraw-Hill Inc; 1991.