Pharmaceutical Research, Vol. 8, No. 4, 1991

Review

Validation of Bioanalytical Methods

H. Thomas Karnes,[1,3] Gerald Shiu,[2] and Vinod P. Shah[2]

Validation of bioanalytical methods used to generate data for pharmacokinetic and bioavailability studies is approached by a variety of techniques and is subject to many different methods of interpretation. This Review puts the various techniques into perspective and discusses pitfalls which may occur in interpretation of validation data. Recovery studies, standardization techniques, and selectivity/specificity are discussed with regard to the intrinsic value of various techniques that are used in validation. Models used for analytical calibration curves are explained in terms of their validity and limitations, along with a presentation of the most common ways to validate the model. Analytical sensitivity and detection limits are presented and discussed with regard to the usefulness of the various definitions. Appropriate means of testing precision and accuracy, the most important factors in assessing method quality, are presented. Stability and ruggedness testing are discussed along with a presentation of ways to assess data acceptability on a daily run basis.

KEY WORDS: quality control; quality assurance; validation; calibration; precision; accuracy.

INTRODUCTION

Quality control procedures in biopharmaceutical analysis associated with drug studies have not been well established. The many different approaches used for evaluation of the quality of analytical methods and results make it difficult to determine whether a prospective analytical method meets the needs of a particular project.

Pharmaceutical science is a dynamic discipline in which today's rules may not fit tomorrow's problems, and it is therefore difficult to establish specific directives regarding method validation. Also, the idea that a minimum set of standards be established is problematic because it is assumed that only the minimum would then be carried out.

The importance of adequate and validated analytical methodology used in biostudies for interpretation of pharmacokinetic data has been discussed earlier (1). This Review puts various bioanalytical validation procedures into perspective.

DEVELOPMENT OF ANALYTICAL METHODS

Method performance is determined primarily by the quality of the procedure itself. The two factors that are most important in determining the quality of a method are selective recovery and standardization.

Analytical recovery of a method refers to whether the analytical method in question provides a response for the entire amount of analyte that is contained in a sample (2). Recovery is usually defined as the percentage of reference material which is measured to that which has been added to a blank. This should not be confused with tests for matrix effects in which recovery is defined as the response measured from the matrix (e.g., plasma) as a percentage of that measured from pure solvent (e.g., water). Results of experiments that compare the matrix to pure solvent are referred to as relative recovery, and true tests of recovery are referred to as absolute recovery.

Absolute recovery is measured as the response of a processed spiked matrix standard expressed as a percentage of the response of pure standard which has not been subjected to sample pretreatment. This can be expressed as a ratio of slopes measured at several concentration points or individual percentages. It should be stated that these are not equivalent because of the effect of weighting with the slopes method.

If an internal standard is used, the recovery of the internal standard should be determined independently. It may be necessary to add an internal standard just prior to injection to compensate for reproducibility problems. Also, in cases where there is no primary standard available for the species that is extracted (e.g., precolumn derivatization), relative recovery may be most appropriate.

Values for recovery not less than 50, 80, and 90% have all been used as numerical acceptance limits. Although it is desirable to attain recovery as close to 100% as possible, it is not needed to provide good accuracy and precision if adequate detection can be attained.

Another important issue at the method development stage is internal versus external standardization. In external standardization, the response of the analyte alone is plotted versus concentration to generate a calibration curve. Inter-

[1] Department of Pharmacy and Pharmaceutics, Virginia Commonwealth University, Box 533, MCV Station, Richmond, Virginia 23298-0533.

[2] Food and Drug Administration, Center for Drug Evaluation and Research, 5600 Fishers Lane, Rockville, Maryland 20857.

[3] To whom correspondence should be addressed.

0724-8741/91/0400-0421$06.50/0

nal standardization requires a functional or isotopic analogue of the analyte which is added to standards and samples prior to sample pretreatment. The response is then generated as a ratio of the signal of the analyte to that of the internal standard (3).

The internal standard technique is very common in bioanalytical methodology especially with chromatographic procedures. The assumption for use of an internal standard is that partition characteristics of the analyte and the internal standard are very similar. This can be a false assumption, and according to Curry and Whelpton (4), the only appropriate uses of nonisotopic analogue internal standards are to serve as qualitative markers, to monitor detector stability, and to correct for errors in dilution and pipetting. Internal standards are usually beneficial for classical instrumentation and manual sample pretreatment. Modern equipment and automation, however, can provide extremely reproducible response measurements. The results of analysis can therefore be adversely affected by the use of an internal standard because of the added variability of the internal standard measurement. This has been experimentally shown through a consideration of the law of propagation of errors (5). The internal standard technique will not inevitably improve, nor will it always adversely affect, the precision of an analytical method. Internal and external standardization techniques may both be evaluated when this issue is in question.

SELECTIVITY/SPECIFICITY

The terms selectivity and specificity are often used interchangeably. The term specific, however, refers to a method which produces a response for only a single analyte. The term selective refers to a method which provides responses for a number of chemical entities which may or may not be distinguished (6). If the response in question is distinguished from all other responses, the method is said to be selective. Since there are very few methods that respond to only one analyte, the term selectivity is usually more appropriate.

In an analytical measurement, where concentration (x) is determined as some function of a response (y), it is desirable that the matrix have no influence on the response. If this condition is met, the analytical method is said to be selective. Frequently, there is some contribution of the matrix which can lead either to a constant or a proportional systematic error. The effects of constant (interference) and proportional (matrix effect) errors on calibration are shown in Figs. 1a and b. Both types of errors are examples of nonspecificity and could occur simultaneously.

In biopharmaceutical analysis interferences are much more problematic. It cannot be assumed that the level of interference in a blank measurement will be equal to that in a measured sample and, therefore, cannot be compensated for by subtraction. In contrast, matrix effects can usually be compensated for by matching the matrix in calibration standards to that in samples.

There is a variety of ways to test selectivity which can provide validation. The simplest test is to demonstrate a lack of response in the blank biologic matrix. In chromatographic analysis, examining blank chromatograms from several sources across the time window of the largest peak can accomplish this (Fig. 2). It is not sufficient to test only one source of blank or to choose one from many that were tested.

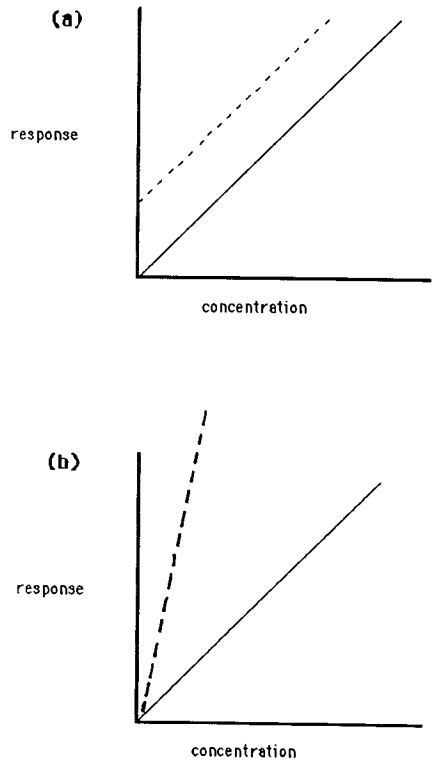

Fig. 1. Calibration curves comparing the "ideal" curve (——) to one showing interference (a; – – –) and to one demonstrating a matrix effect (b; – – –).

Another approach is to test whether the intercept of the calibration curve is significantly different from zero. This is done with a one-sided t test (7) and provides a quantitative assessment. This test can be very liberal or rigid, depending on the reproducibility of calibration.

For competitive binding assays, lack of interference is generally indicated by a lack of measurable response in the blank along with an evaluation of cross-reactivity to structurally related or concomitant substances. Lack of cross-reactivity is generally accepted to be no response when the

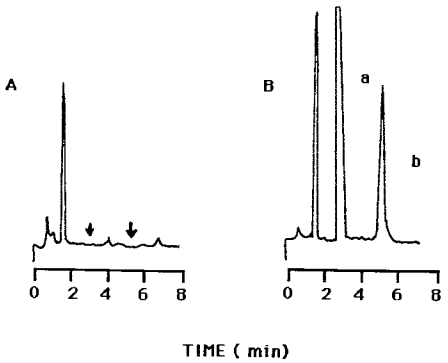

Fig. 2. Chromatograms showing the absence of interfering peaks (indicated by arrows) (A) over the time window of elution for the highest concentration standard (B).

substance is present at 1000 times the lower limit of quantitation for the analyte or ≤0.1% cross-reactivity.

The presence of matrix effects can be tested either by a comparison of slopes between matrix standards and nonmatrix standards or by standard addition. In standard addition, known quantities of analyte are added to a real sample which may contain metabolites. The slope of the measured responses divided by the y intercept should equal the x intercept within experimental error (8). This technique can be useful for demonstrating a lack of matrix effect with regard to metabolites when primary standards of the metabolites are not available.

CALIBRATION

The calibration curve should be constructed using at least five to seven values from the expected range of concentrations. Although some analytical procedures may require nonlinear calibration, it is conventional to use a linear model and univariant regression. In this model, the independent variable (x) is concentration and the dependent variable (y) is response. This distinction is necessary since univariant regression minimizes residuals around y and assumes x to be errorless, and it is more appropriate to assume negligible error associated with the concentration axis. Other assumptions involved in linear regression analysis are that the calculated residuals are independent, are normally distributed, and have equal variances. The condition of equal variances is termed homocedasticity and, unlike the first two conditions, is frequently not met for analytical data. A test of homocedasticity can be carried out by observing a plot of residuals versus concentration (Figs. 3a and b). The most common occurrence of nonhomocedasticity or heterocedasticity is an increase in variance as a function of concentration (Fig. 3b). In such cases, the assumption of homocedasticity is not justified and a weighted regression may be performed (9).

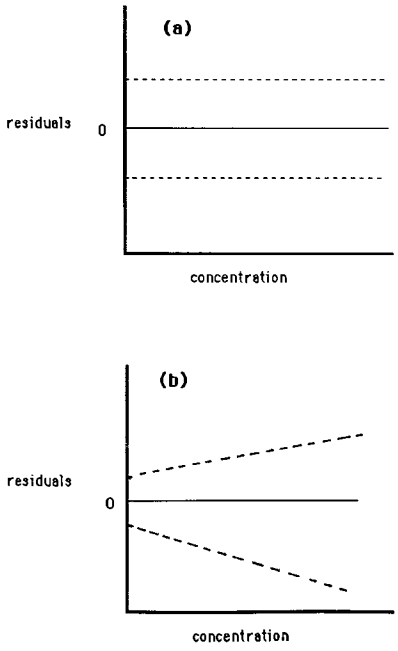

Fig. 3. Residuals versus concentration plots (– – –) showing a constant error term (a) and an error term which increases with concentration (b).

The most appropriate weighting factor is the inverse of the variance of the standard point, although $1/x$, $1/y$, and $1/x^2$ (x = concentration; y = response) are valid approximations of this variance (10). The effect of weighting the calibration curve is shown in Table I. There is usually some sacrifice in the accuracy of reverse predicted standards at the high end of the range for improved accuracy at the low end.

Data that should be linear by theory but provide a better fit by a nonlinear function suggest a problem in the system. A better fit may also be provided by a nonlinear model in cases where the condition of homocedasticity is not met. Linearity can be demonstrated over a given range by simple observation of a calibration plot. If there is no deviation from linearity at either of the extremes of the curve, the calibration range can be considered appropriate. A generally accepted criterion for the upper or lower boundary is the point where the slope of the line deviates from the overall slope by not more than 5% (11).

The quality of fit can be evaluated using tables that list reverse calculated standard points as compared to the nominal value; tables that list correlation coefficients, slopes, and intercepts; lack-of-fit analysis; and log concentration versus log response plots. The acceptability of reverse calculated mean and individual data should be in line with the acceptance criteria set for the evaluation of quality control samples. The criteria for correlation coefficients, slopes, and intercepts should vary according to the method. The downfall of lack-of-fit analysis for analytical data is that the more precise the data, the less the likelihood of passing the test.

A plot of log concentration versus log response should ideally provide a slope equal to one for a linear model (11). The closeness of the slope to one may be used as a criterion of acceptance for linear data. Although partition processes should theoretically be linear, a certain degree of nonlinearity may be acceptable. This is handled best by the almost-linear approach in which a fractional exponent is used to impart some curvature to calibration (11). The degree to which the exponent can deviate from unity is controlled to be within acceptance limits for linearity.

SENSITIVITY AND DETECTION LIMIT

There is a great deal of confusion over the terms related

Table I. Comparison of Reverse Predicted Standard Points Using Weighted and Unweighted Regression

	Average deviation	
Concentration (ng/ml)	Unweighted (%)	Weighted 1/conc. (%)
20	49.1	9.4
50	18.9	8.8
100	18.5	9.5
500	3.1	4.1
1000	7.9	7.1
2000	2.2	3.9

to the ability to assay low concentrations. The term most frequently used is sensitivity (S). A method is said to be sensitive if small changes in concentration cause large changes in analytical response (12). Although sensitivity (defined as the slope of the calibration curve) is very closely related to the ability of a method to measure small concentrations, the term has no allowance for reproducibility of the measurement and cannot be used appropriately as a criterion to describe this concept. A term which recognizes that lower variability of a measurement can be just as important to the ability to detect small concentrations as the slope of the curve is S/s_b, where s_b represents the standard deviation of the blank measurement (13). This term has not been generally accepted, however, and the ability to detect small concentrations can be expressed as the limit of detection (LOD) (14), limit of quantitation (LOQ) (14), or limit of guarantee of purity (LOGP) (16). The limit of detection is calculated as $\text{LOD} = 3s_b/S$, where 3 is a factor for a 99.9% level of confidence. This value represents the smallest concentration that can be distinguished from the blank. While this may be an adequate measure of the theoretical limit of an analytical method, concentrations measured at this level would be indistinguishable from zero measurements by a large probability.

The LOGP is a value that represents the concentration at which a sample can be distinguished from blank with greater probability and can be calculated as $\text{LOGP} = 6s_b/S$, where s_b represents the standard deviation of the blank and is assumed to be equal to the standard deviation of a sample measurement at LOGP. This value is used to identify whether an analyte is present or absent in a sample but does not indicate a value above which reliable quantitation can be accomplished.

The limit associated with reliable quantitation is the LOQ and it is defined as the concentration above which quantitation is possible within a certain preset level of certainty. Although this value cannot be accurately calculated a priori, it is estimated to be $\text{LOQ} = Ks_b/S$, where K is a factor indicating the desired precision at the lower limit (i.e., 10 for 10% RSD, 20 for 5% RSD, etc.). The RSDs that are deemed acceptable may vary but usually range from 10 to 20%. Concentrations below LOQ should not be quantitatively reported but may be reported as simply "present" or as semiquantitative numbers. They should not be used in the interpretation of results without appropriate weighting factors. Practical validation is accomplished by defining the LOQ as the concentration of the lowest standard and demonstrating the appropriate level of certainty in reverse calculated standards. The lowest standard then becomes the operational limit since one should not extrapolate standard curve data beyond the range of standardization. Although the practice is widely accepted, it should be pointed out that the standard had influence on the regression line from which it was calculated and does not represent the uncertainty of an unknown sample which is measured at that same concentration.

PRECISION AND ACCURACY

Precision and accuracy together determine the error of an analytical measurement and are the primary criteria used when one judges the "quality" of an analytical method. Accuracy or method bias is generally recognized as the difference between the mean of a set of results and the "true value." Precision which refers to the variability of measurements within a set is often confused with reproducibility and repeatability. Reproducibility should be used to describe the closeness of agreement between results obtained with the same method under different conditions, whereas repeatability refers to agreement between successive measurements on the same sample.

Precision and accuracy are considered together because they are interdependent in assessing the acceptability of a method. For example, the term accuracy applied to a single quality control measurement is inappropriate because a single measurement has incorporated into it both random and systematic error. If one considers the random error to be the width of the Gaussian distribution at a specified confidence interval of 99% (estimated by $\pm 2.58 \times$ the standard deviation; SD), the accuracy of the method can be obtained as the difference between the measured mean (x) and the true value (μ). The total error (E) which applies to an individual value can therefore be calculated as follows (17):

$$|x - \mu| + 2.58\ \text{SD} = E$$

An assessment of accuracy must therefore be carried out on mean values, and if acceptance limits are to be applied to single quality control values, the range must be wide enough to include both the random and systematic error.

The "true value" for accuracy assessment can be obtained in two ways. The matrix of interest can be spiked in bulk at a known concentration of the analyte and the true value is taken as this concentration. One can also compare results of a method with results of an established "reference method." The reference method approach can be used with real samples that would contain metabolic products. This approach, however, assumes no systematic error in the reference method, which is generally a false assumption.

It may be helpful to determine whether the bias is due to random error alone. This is done with a t test to determine whether the mean value differs significantly from the true value. If the difference is significant, the deviation between the mean of the measured results and the true value is an estimate of the bias. Significance testing is generally not carried out and the accuracy of a method is usually expressed simply as the percentage difference of the mean from the true value.

There are two basic questions that should be addressed by the evaluation of precision. The first is a determination of whether the analytical method is precise enough for a particular study. This assessment is usually carried out prior to the analysis of samples. The second is to monitor the quality of a method during sample analysis and is intended to detect transient problems.

Precision is usually assessed on both a within-batch and a between-batch basis. Batch terminology is more appropriate than "within day" and "between day" since between-batch assessment is not always carried out with a single batch per day and some batches may be of sufficient size that more than 1 day is required for analysis of a batch. Within-batch assessment should be considered as a measure of the precision of a method under optimal conditions. The be-

tween-batch precision is considered to be a better representation of the precision one might observe during routine conduct of a method because these data are generally subjected to a greater number of sources of variability.

Precision is generally assessed by repeated analysis of known quality control samples measured independent of the calibration curve and spiked at concentrations to represent the low, medium, and high ranges of calibration. All quality control results should be considered for method assessment unless contraindicated by poor chromatography or equipment failure. The goal of the "during-study" assessment, however, is to monitor the method. For quality control during sample analysis, values that fall outside of a set rejection range should be deemed "out of control" and samples corresponding to these controls should be reanalyzed. Since this assessment is the indicator of method function, one should not report values during a time in which the method was not functioning according to the quality control data.

THE "BATCH" CONCEPT AND ACCEPTANCE OF DATA

There is general agreement that quality control (QC) samples should be placed throughout a batch of samples during operation of an analytical procedure. There are different techniques applied, however, with regard to the number of QC samples in a batch, the sequence of QC placement, the order in which QC concentrations are run, and the criteria for acceptance of sample data based on QC. The number of QC samples per batch should depend on the size of the batch and the stability of the analytical method. It is frequently difficult to determine the stability of a method a priori and it is advisable to let the size of the batch determine the number of QC samples to be run. It is also prudent to disperse QC samples evenly in a low–high, high–low sequence throughout the batch. In this way, maximum detectability of analytical problems can be achieved. Prior to sample analysis when the method is evaluated it is a good idea to simulate batch conditions with QC samples to detect problems of drift. If controls are arranged low to high, then high to low, carryover problems can also be detected.

There are generally two ways used to establish acceptability of samples within a run. The bracket approach establishes that acceptable sample results should be taken only from between acceptable QC samples. Although this approach is appropriate, an additional criterion should be established to detect concentration-dependent problems. The second method is the application of the acceptance criterion on the entire batch. This criterion is usually set to allow a certain percentage (usually not more than 33.3%) of unacceptable quality control samples, along with some criteria for concentration-dependent problems, before an entire run is judged unacceptable. In this method, however, transient problems that may occur during an analytical run are not compensated for and an entire run may be rejected, although a large portion of the run may have occurred in the absence of the problem.

Criteria of acceptance for QC samples can be established with a fixed-range approach or a confidence-interval approach. The fixed-range approach is based on a combined accuracy and precision criterion with an arbitrary range around the "true value." This range can be established anywhere between 10 and 25%. The advantage of a fixed-range approach is that all methods must yield values of a certain minimum quality to be acceptable. A quality control program, however, not only should assess the quality of a method but should ensure that the method is operating acceptably on a routine basis. The confidence interval approach does this through establishment of a range around the mean value for QC samples. This range is based on a factor, chosen to represent the level of confidence, times the standard deviation. In this way, the precision of the method itself determines the acceptance range and wide ranges would be allowed for relatively imprecise methods. In contrast, the fixed-range criteria may be too wide for methods that are relatively precise and too narrow for methods that are imprecise as is illustrated in Fig. 4. All methods do not have equivalent precision and the decision regarding adequate precision should be made independent of the decision on whether the method is operating as expected. To address the quality of the method with a confidence interval approach, absolute criteria can be applied to overall relative standard deviations and separate accuracy criteria can be applied to mean QC values. This will prevent acceptance of data demonstrating a significant systematic error.

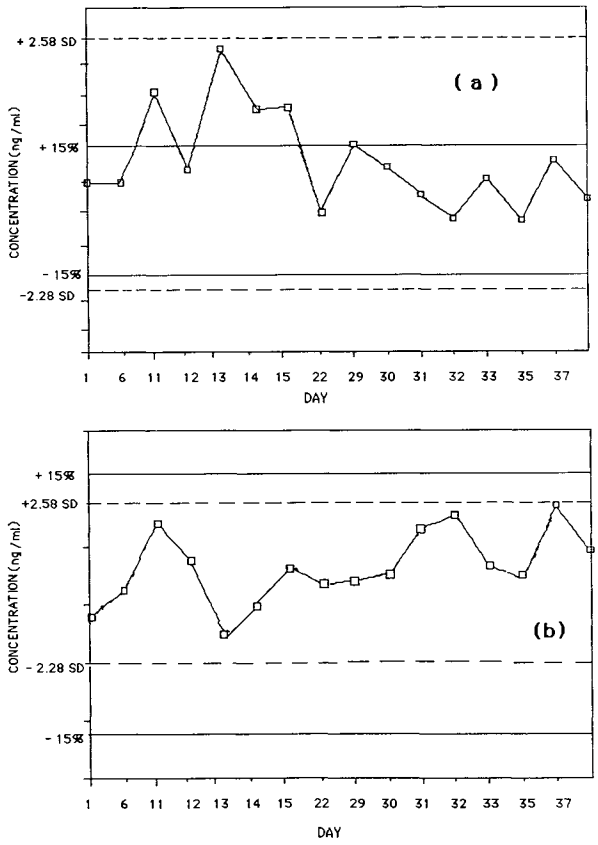

Fig. 4. Quality control charts showing confidence-interval (±2.58 Sb; – – –) and fixed-range (±15%; ——) approaches to evaluation of quality control data. Examples show situations where the fixed-range approach is relatively narrow (a) and where the fixed-range approach is relatively wide (b).

SAMPLE STABILITY

In situations where samples are stored, the stability of drugs during the storage period should be ensured. Assessment of drug stability can be accomplished by analyzing spiked samples immediately and on subsequent days for the anticipated storage period. When the integrity of the drug is affected by freezing and thawing, spiked samples should be stored in individual containers and appropriate caution should be employed for study samples. Instability due to freezing and thawing should be tested independently when study samples are to be repeated after thawing and refreezing. In the event of indicated instability, appropriate additives (buffers, antioxidants, or enzyme inhibitors) may be essential to minimize degradation of the analytes.

RUGGEDNESS

The ruggedness of a method refers to the reproducibility of the method under different conditions. This can be assessed experimentally, to determine the effect of a changing environment on a method, usually through factorial approaches (18). In a method validation sense, the goal is usually to determine a priori whether a different set of conditions will have an effect on a method. This is done through a process of cross validation in which the new conditions are applied and the data are evaluated to determine whether the new conditions had an effect. Typical reasons for cross validation would include transfer of the method from one analyst to another, significant instrumental or procedural modifications, and a significant time lapse between periods of operation. Cross validation will involve the regeneration of validation data under the new conditions and comparison of these results to the original.

Another way to address ruggedness testing is to observe QC results over time or under different conditions to determine if results have changed with time or environment. The easiest way to monitor results over time is to construct Levy–Jennings charts (19) and visually determine if trends or shifts in the data can be observed. These data can be statistically analyzed to determine a criterion for acceptability (20). Assessment of ruggedness under different conditions a posteriori is necessary when one wishes to carry out interlaboratory comparisons. This is frequently done in biopharmaceutical analysis with multisite studies and the laboratories in question need not use the same methodology. Results of such a comparison would be analyzed appropriately with a one-way analysis of variance (21).

CONCLUSION

One of the major causes of deficient biopharmaceutics submissions to the FDA has been inadequate analytical method validation (1). The essential performance characteristics for analytical methodology have been discussed here in detail to provide guidance to bioanalytical chemists.

REFERENCES

1. V. P. Shah. Analytical methods used in bioavailability studies: A regulatory viewpoint. *Clin. Res. Pract. Drug Reg. Affairs* 5:51–60 (1987).
2. J. K. Taylor. *Quality Assurance of Chemical Measurements,* Lewis, Chelsea, Mich., 1987.
3. R. V. Smith and J. T. Stewart. *Textbook of Biopharmaceutic Analysis,* Lea and Febiger, Philadelphia, 1981.
4. S. H. Curry and R. Whelpton. Statistics of drug analysis, and the role of internal standards. In E. Reid (ed.), *Blood Drugs and Other Analytical Challenges,* Ellis Horwood, Chichester, 1978, pp. 29–41.
5. P. Haefelfinger. Limits of the internal standard technique in chromatography. *J. Chromatogr.* 218:73–81 (1981).
6. D. L. Massart, B. G. M. Andeginste, S. N. Deming, Y. Michotte, and L. Kaufman. *Chemometrics a Textbook,* Elsevier, New York, 1988.
7. R. L. Anderson. *Practical Statistics for Analytical Chemists,* Van Nostrand, New York, 1987.
8. B. E. Cooper. *Statistics for Experiments,* Pergamon Press, Oxford, 1975.
9. E. L. Johnson, D. L. Reynolds, D. S. Wright, and L. A. Pachla. Biological sample preparation and data reduction concepts in pharmaceutical analysis. *J. Chromatogr. Sci.* 26:372–379 (1988).
10. H. G. Boxenbaum, S. Riegelman, and R. M. Elashoff. Statistical estimations in pharmacokinetics. *J. Pharmacokin. Biopharm.* 2:123–148 (1974).
11. C. A. Dorschel, J. L. Ekmanis, J. E. Obertholtzer, F. V. Warren, Jr., and B. A. Bidlingmeyer. LC detectors: Evaluation and practical implications of linearity. *Anal. Chem.* 61:951A–968A (1989).
12. International Union of Pure and Applied Chemistry. Nomenclature, symbols, units and their usage in spectrochemical analysis, II. *Anal. Chem.* 48:2294–2296 (1976).
13. N. E. Saris. International Federation of Clinical Chemistry—Committee on Standards. Quality control in clinical chemistry, I. *J. Clin. Chem. Clin. Biochem.* 18:69–77 (1980).
14. G. L. Long and J. D. Winefordner. Limit of detection—A closer look at the IUPAC definition. *Anal. Chem.* 55:712A–722A (1983).
15. American Chemical Society's Committee on Environmental Improvement. Guidelines for data acquisition and data quality evaluation in environmental chemistry. *Anal. Chem.* 52:2242–2249 (1980).
16. H. Kaiser. *Two Papers on the Limit of Detection of a Complete Analytical Procedure,* Hafner, New York, 1969.
17. J. O. Westgard and P. L. Barry. *Cost-Effective Quality Control: Managing the Quality and Productivity of Analytical Processes,* AACC Press, Washington, D.C., 1986.
18. W. J. Youden and E. H. Steiner. *Statistical Manual of the A.O.A.C.,* Association of Official Analytical Chemists, Washington, D.C., 1975.
19. S. Levey and E. R. Jennings. The use of control charts in the clinical laboratory. *Am. J. Clin. Pathol.* 20:1059–1066 (1950).
20. G. S. Cembrowski, J. O. Westgard, A. A. Eggert, and E. C. Toren. Trend detection in control data: Optimization and interpretation of Trigg's technique for trend analysis. *Clin. Chem.* 21:1396–1405 (1975).
21. R. L. Anderson. *Practical Statistics for Analytical Chemists,* Van Nostrand, New York, 1987.