Pharmaceutical Research, Vol. 10, No. 10, 1993

Research Article

Precision, Accuracy, and Data Acceptance Criteria in Biopharmaceutical Analysis

H. Thomas Karnes[1,2] and Clark March[1]

Received October 16, 1992; accepted April 6, 1993

Accuracy and precision are the most important criteria in the assessment of an analytical method, and monitoring quality control during sample analysis is essential to ensure the validity of reported results. Various approaches to testing accuracy, precision, and quality control were applied to 10 analytes from seven chromatographic bioanalytical methods. These methods include fixed interval bias and significance testing for accuracy; fixed interval percentage relative standard deviation (%RSD) and analysis of variance (ANOVA) approaches for precision; ±20% fixed range, 99% confidence interval, multiple rules, and range chart for individuals approaches for quality control acceptance criteria. Quality control approaches were also applied to the entire run and to a bracketed approach whereby results are considered valid only if bracketed by acceptable quality control. Accuracy and precision were assessed for six runs of each analyte at three concentrations established to represent the calibration range of the analytical method. Quality control acceptance criteria were evaluated using all data sets from each of the analytical methods collected during the course of running various numbers of real samples. The data suggest that the fixed interval bias criteria for accuracy was a more liberal method of accuracy assessment because three of the seven methods would have been rejected according to the significance testing criteria whereas all were acceptable by the fixed internal bias criteria. Precision can be effectively assessed for between- and within-run data by criteria set on unconfounded %RSD values or by separation of the sources of variation using an ANOVA approach applied to confounded data. The percentage of samples rejected for the 99% confidence interval applied to brackets, the multiple rules approach applied to the entire run, and the individual range chart approach applied to brackets were comparable and were found to be 7.0, 6.2 and 8.3 percent respectively. The ±20% fixed range criteria applied to the two-thirds of the run resulted in just 2.9% sample rejection and was not considered comparable to the other methods.

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

INTRODUCTION

There has been much controversy in recent years over which approaches should be used to assess bioanalytical methods with regard to their precision and accuracy. Precision and accuracy are the most important criteria in determining whether a method is suited to a particular task (validation) and whether data generated under routine use of a bioanalytical method are acceptable (acceptance criteria). This controversy was recently addressed in guidelines issued by a joint conference of the FDA, AAPS, AOAC, HPB, and FIP (1). The authors of the position paper made specific recommendations for validation of accuracy and precision and recommended a fixed range approach to determine the acceptance of data. The conference report recommended a second criterion involving accuracy and precision which allows for 15% accuracy and precision above the limit of quantitation (LOQ) and 20% at or near the LOQ. It was not clear, however, whether this criterion was to be applied only to data that passed the ±20% fixed range criterion or to all data. It was also recommended that this second criterion be "provided" for intraday and interday experiments which confuse the goals of validation and acceptance of runs. The most logical interpretation of this recommendation is that the 15 and 20% accuracy and precision criteria be applied to the mean data from all analytical runs from a particular study, because the number of recommended data points of each control concentration, two, is insufficient to evaluate accuracy and precision on a single run or "intraday" basis. This would then be a criterion applied to data from an entire study and it is not clear what course should be taken in the event of failure with regard to this criterion since the recommendation is to "provide" the assessments. A "control chart approach" was recommended as an acceptable alternative to the fixed range criterion for data acceptance, which suggests that some controversy may still exist.

It has been shown that a fixed range approach may be too conservative in some instances and too liberal in others for the purpose of determining whether an analytical method is working as it should (2). The fixed range approach also

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

[2] To whom correspondence should be addressed.

0724-8741/93/1000-1420$07.00/0 © 1993 Plenum Publishing Corporation

confounds the issues of precision and accuracy into one criteria. This may be undesirable because problems with accuracy (systemic bias) often can and should be corrected (3). The total error component due to imprecision (random error) is generally unavoidable but should be monitored to ensure that precision does not degrade from established acceptance criteria. The fixed range approach has been justified in statistical terms for a ±10% range (4) but acceptance of this method requires that more than one-third of the control samples in an analytical run be outside the range before a run is rejected.

Criticisms of the "control chart" or confidence interval approach include the absence of an accuracy criterion, skepticism regarding the detection of errors, and a high probability of false rejection, leading to cost inefficiency of the process. The accuracy criterion can be addressed simply by imposing a criterion for accuracy in addition to the confidence interval. It should be mentioned that the accuracy criterion should be applied to mean rather than individual data so that the systematic error component may be reduced and pure assessment of accuracy can be obtained (3). There are several publications which address theoretical probabilities of error detection and false rejection for a variety of quality control methods involving confidence intervals (5–8). In an experimental sense it is difficult to determine these probabilities accurately since it is not known conclusively when a "true" error occurs.

In this paper we have evaluated real data for 10 different analytes in terms of precision, accuracy, and acceptance criteria. The evaluation consisted of the application of various validation and quality control approaches to the data and comparing each in terms of rejection rates.

METHODS

Ten analytes from seven chromatographic bioanalytical methods were selected to represent various separation, detection, and extraction modes. These methods are characterized in Table I and were carried out by BioClin Analytical at the Virginia Commonwealth University. Calibration of all methods was accomplished with a minimum of six matrix matched standards at different concentrations. Each method utilized three levels of quality control samples, the lowest of which fell between the limit of quantitation (2) and the next-highest standard concentration point. The two additional controls were selected to represent the mid and upper range of calibration for the method. The specific validation and control procedures used are described below.

Accuracy in Validation

Percentage Difference from Actual

One datum point for each control was taken from six different runs performed over several days. The means of the six results were calculated (Quattro Pro, Version 3.0, Borland International, Inc., Scotts Valley, CA) and compared to the spiked value to determine the percentage difference from actual (%DFA) according to the formula

$$\%\text{DFA} = [(\text{mean-spiked})/\text{spiked}]100$$

The results were tabulated for each analyte by control concentration.

t *Test*

The same data used to calculate %DFAs were subjected to a two-sided *t* test for determining a significant difference between the mean of the data and the spiked value with a 95% level of confidence. Quattro files were imported into Statgraphics, Version 5 (STSC, Inc., Rockville, MD), and the *t* test was performed by the Statgraphics software. The results were tabulated by control concentration for each analyte.

Precision in Validation

Between-Run and Within-Run Precision

The same data used in the accuracy determinations were used for the calculation of the between-run percentage relative standard deviation (%RSD). The within-run %RSD resulted from analysis of six controls at each concentration for each analyte within one analytical run. The calculations were performed using the sample standard deviation function of Quattro Pro and calculating

$$\%\text{RSD} = (\text{standard deviation}/\text{mean})100$$

Table I. Characteristics of Analytical Methods Used for Assessments

Analyte	Compound type	Extraction	Separation	Detection	LOQ[a]
A	Sulfonylurea, 2° amine,	Double SPE[b]	HPLC	UV	10 ng/mL
B	Sulfonylurea, 2° amine	Double SPE	HPLC	UV	5 ng/mL
C	Sulfonylurea, 2° amine, alcohol	Double SPE	HPLC	UV	5 ng/mL
D	Sulfonylurea, 2° amine, carboxylic acid	Single SPE	HPLC	UV	10 ng/mL
E	2° amine, alcohol	Double LLE[c]	HPLC	FL	1 ng/mL
F	Carboxylic acid	Protein precip.	HPLC	UV	0.5 μg/mL
G	Carboxylic acid, 2° amine	Single LLE	HPLC	UV	5 ng/mL
H	2° amine	Triple LLE	GC	ECD	5 ng/mL
I	1° amine	Triple LLE	GC	ECD	5 ng/mL
J	Carboxylic acid, 2° amine	Single SPE	HPLC	FL	50 ng/mL

[a] Limit of quantitation.
[b] Solid phase extraction.
[c] Liquid–liquid extraction.

These results were tabulated for each control concentration by analyte.

ANOVA and Confounded Precision

Data from six analytical runs with two values for each control concentration were collected. The results were analyzed using a one-way analysis of variance with the Statgraphics software. The between-run and within-run mean square of variance was used to calculate %RSD manually as follows:

$$\%\mathrm{RSD} = [(\text{mean square variance})^{0.5}/\text{mean}]100$$

All 12 data points were also used to determine a confounded %RSD. The results were tabulated for each analyte at each control concentration.

Data Acceptance Criteria

99% Confidence Interval

All control values for each analyte were tabulated in Quattro Pro and a mean and sample standard deviation were calculated. Precision acceptance criteria were determined using a range of mean ± 2.58 × standard deviation. If a control value was outside this range, it was rejected and the number of rejected samples was counted using one of two approaches. The first approach (referred to as Two-Thirds of Run) was as suggested by the conference report (1), wherein two-thirds of the controls being acceptable allows acceptance of the entire run. The second approach (referred to as Bracketed) allowed acceptance of only those results that were bracketed before and after by acceptable control values.

The same data were also reviewed using a two-step evaluation of accuracy then precision. Each control result was tested for inclusion in the range of mean ± 25% as an accuracy measure. If a result was outside this range, it was excluded from the determination of the mean and standard deviation. The precision range was then determined as mean ± 2.58 × standard deviation. Any values outside this range were also rejected and repeat samples were counted based upon those controls rejected for accuracy or precision. The Two-Thirds of Run and Bracketed approaches were both used in counting the number of rejected samples.

20% Fixed Range

The control values were evaluated according to the recommendations of the Conference Report; acceptable controls are within the range of spiked value ±20%. Once rejected results were determined, the number of repeat samples was counted using the Two-Thirds of Run and Bracketed approaches. The second criterion recommended by the conference report was interpreted to be a criterion applied to an entire set of runs rather than a single run and was therefore not applied as a run acceptance criterion.

Westgard Multirules

The six-rule scheme was used that consisted of the 1_{2s}, 1_{3s}, 2_{2s}, R_{4s}, 4_{1s}, and $10\bar{x}$ rules applied in order only if the first rule failed (9). The number of repeat samples was counted based upon Westgard's approach of rejecting the entire run or the Bracketed approach.

Individual Range Charts

Using Quattro Pro, recovery, average recovery, range, and moving range calculations were performed as outlined by Bolton and Lang (10). Acceptable ranges for recovery and range (reproducibility) were calculated and applied as acceptance criteria for controls. The number of rejected controls was determined by using only the recovery criterion and by using the recovery and range criteria. The number of repeat samples was counted using both the Two-Thirds of Run and the Bracketed approaches.

RESULTS AND DISCUSSION

Validation of Accuracy and Precision

Issues of interest in the validation stage of analytical method assessment are whether the method is accurate and precise enough to be used for a specific purpose. This purpose is determined by other factors which impact the quality of an analytical method and are validated separately (2,11). In the case of biopharmaceutical analysis this usually relates

Table II. Accuracy Assessments Comparing Difference Criteria to *t* Test for Acceptability

Analyte	Concentration (ng/mL)	%DFA	*t* test (*P*)
A	2500	−0.95	0.65
	250	−6.06	0.067
	15	−1.60	0.63
B	125	1.70	0.47
	40	5.40	0.003[a]
	15	6.40	0.19
C	400	−1.66	0.43
	40	3.52	0.22
	15	5.27	0.26
D	750	1.89	0.030[a]
	150	−5.15	0.036[a]
	15	−5.27	0.13
E	75	−2.13	0.42
	25	−2.68	0.15
	5	−8.80	0.11
F	125	0.47	0.58
	20	−0.35	0.81
	1	−2.00	0.60
G	700	−1.89	0.66
	150	−1.81	0.62
	15	0.73	0.89
H	100	−4.01	0.0001[a]
	35	0.94	0.18
	7.5	−4.40	0.023[a]
I	100	0.89	0.83
	35	−8.77	0.072
	7.5	−3.47	0.27
J	700	−1.92	0.24
	300	0.88	0.35
	75	0.52	0.76

[a] Rejected value.

to pharmacokinetic studies. The extent to which bias and imprecision are allowed is somewhat subjective and should be guided by the requirements of the decision to be made with the data. Although using set criteria in all cases has merit from a regulatory point of view, not all questions need be addressed with the same level of accuracy and precision. Using set criteria in all cases does not make sense in terms of efficiency, especially if criteria are set for the most demanding of pharmacokinetic questions. It also is not prudent to establish criteria based on the least demanding questions because more demanding questions could not be adequately answered.

Accuracy

The most common way to assess the accuracy of a method is to set an acceptance criterion for %DFA (usually 10 or 15%) on mean analytical data using the spiked concentration of the control as the actual value. An alternate way is to determine whether the assayed mean value is statistically different from the actual value using a *t* test at 95% confidence. The results for comparison of these two methods are shown in Table II for each of the bioanalytical procedures considered. All values for %DFA are within 10%, whereas five controls representing three analytes would be judged unacceptable according to the *t*-test criterion. Although the *t* test would appear to possess statistical justification, it is interesting to note that the higher concentrations, which are more precise, are rejected more often than lower concentrations by the *t*-test criterion. This suggests a problem with *t*-test evaluation of accuracy since there is such a large dependence on the precision of the method and even a very small bias will be detected in precise data.

Precision

Validation of precision is often carried out by establishment of a criterion on relative standard deviation (RSD). This criterion is generally established between 10 and 20% depending on the method and the requirements of the results. RSD is usually evaluated on both a between-run and a within-run basis. Between-run data are considered to be the precision of the method under "real-world" conditions and the within-run data are considered to be precision under "ideal conditions." This is evaluated by analysis of single results over many runs for between-run and multiple results during a single run for within run. If data are collected with multiple results per run, an alternate technique is to apply a one-way analysis of variance (ANOVA) which separates out the sources of variance due to within- and between-run fac-

Table III. Precision Measured as Percentage Relative Standard Deviation (%RSD) Comparing Between- and Within-Run Data Confounded, Not Confounded, and Analyzed with Analysis of Variance ANOVA

Analyte	Concentration (ng/mL)	%RSD (n = 6)		ANOVA		
		Between run	Within run	Between run	Within run	Confounded
A	2500	4.93	2.63	3.19	4.86	4.18
	250	6.77	3.50	5.73	11.95	9.63
	15	7.81	8.07	9.08	7.87	8.44
B	125	5.24	2.59	5.86	3.41	4.69
	40	2.32	2.48	4.48	2.89	3.7
	15	9.61	8.25	11.61	7.49	9.87
C	400	4.80	2.31	5.00	2.75	3.94
	40	5.89	4.96	5.57	7.99	6.99
	15	9.52	5.73	4.59	9.79	7.86
D	750	1.51	3.39	4.23	5.29	4.84
	150	4.67	2.76	5.19	4.51	4.83
	15	7.47	6.03	6.12	6.88	6.55
E	75	6.11	7.82	9.95	2.66	6.99
	25	3.94	3.01	6.40	3.79	5.14
	5	12.39	3.10	18.15[a]	2.48	12.37
F	125	1.92	0.55	2.72	0.96	1.97
	20	3.37	2.67	4.52	2.61	3.61
	1	11.33	7.45	12.00	8.69	10.3
G	700	9.99	8.32	7.75	8.41	8.11
	150	8.46	4.55	8.42	7.90	8.14
	15	11.77	10.09	16.88[a]	10.18	13.64
H	100	0.91	3.05	2.79	2.06	2.42
	35	1.49	1.94	1.65	1.78	1.72
	7.5	3.46	2.73	2.85	2.26	2.54
I	100	9.89	6.13	11.10	5.12	8.38
	35	10.33	7.33	7.82	7.43	7.61
	7.5	6.95	9.43	10.08	6.59	8.36
J	700	3.59	3.62	5.63	1.57	3.97
	300	2.07	2.40	3.76	0.95	2.63
	75	3.95	4.36	6.27	3.21	4.85

[a] Rejected value.

Table IV. Criteria for Acceptance of Data Comparison of Confidence Interval, Fixed Range, Multiple Rules, and Range Chart Approaches

Analyte	*n* (controls)	Number of controls rejected	Number of samples repeated (% repeated)	
			Run brackets	Two-thirds of run
(A) 99% confidence interval considering only precision				
A	66	0	0 (0)	0 (0)
B	66	2	40 (7.3)	50 (9.1)
C	66	2	50 (9.1)	0 (0)
D	276	4	70 (3.0)	0 (0)
E	259	1	20 (0.8)	0 (0)
F	90	1	20 (2.5)	0 (0)
G	162	2	170 (11.8)	290 (20.1)
H	120	3	50 (4.1)	0 (0)
I	120	1	10 (0.8)	0 (0)
J	78	0	0 (0)	0 (0)
Total	1303	14	430 (3.7)	340 (2.9)
(B) Sequential application of accuracy criterion (mean ± 25%) and 99% confidence interval				
A	66	3A,0P[a]	50 (9.1)	0 (0)
B	66	1A,0P	30 (5.5)	0 (0)
C	66	3A,0P	70 (12.7)	0 (0)
D	276	3A,2P	90 (3.9)	0 (0)
E	259	8A,0P	140 (5.9)	0 (0)
F	90	0A,1P	20 (2.5)	0 (0)
G	162	11A,0P	290 (20.1)	290 (20.1)
H	120	0A,3P	50 (4.1)	0 (0)
I	120	5A,0P	80 (6.6)	0 (0)
J	78	0	0 (0)	0 (0)
Total	1303	40	820 (7.0)	290 (2.5)
(C) Fixed range accuracy only (±20%)				
A	66	4	60 (10.9)	0 (0)
B	66	8	130 (23.6)	50 (14.3)
C	66	5	100 (18.2)	0 (0)
D	276	4	60 (2.6)	0 (0)
E	259	17	260 (10.9)	0 (0)
F	90	0	0 (0)	0 (0)
G	162	24	560 (38.9)	290 (20.1)
H	120	0	0 (0)	0 (0)
I	120	7	110 (9.0)	0 (0)
J	78	0	0 (0)	0 (0)
Total	1303	69	1280 (10.9)	340 (2.9)

(D) Westgard rules acceptance criterion—applied in order

Analyte	Type and (number) of failures	Number of samples repeated (% rejected)		
		Run brackets	Entire run	Two-thirds of run
A	2_{2s} rule (2)	70 (12.7)	100 (18.2)	50 (9.1)
B	1_{2s} rule (1)	30 (5.5)	50 (9.1)	0 (0)
C	None	0 (0)	0 (0)	0 (0)
D	2_{2s} rule (2)	30 (1.3)	100 (4.3)	0 (0)
E	R_{4s} rule (3)	110 (4.6)	240 (10.0)	0 (0)
F	None	0 (0)	0 (0)	0 (0)
G	R_{4s} rule (1)	40 (2.8)	80 (5.5)	0 (0)
H	2_{2s} rule (1) and R_{4s} rule (1)	50 (4.1)	100 (8.2)	0 (0)
I	2_{2s} rule (1)	30 (2.5)	50 (4.1)	0 (0)
J	None	0 (0)	0 (0)	0 (0)
Total	12	360 (3.1)	720 (6.2)	50 (0.4)

Table IV. Continued

(E) Range chart for individuals approach (numbers in parentheses represent failures based on repeatability)

			Number of samples repeated (% rejected)		
Analyte	*n* (controls)	Number of controls rejected (repeatability failures)	Run brackets	Entire run	Two-thirds of run
A	66	4	80 (14.5)	150 (27.3)	50 (9.0)
B	66	3	60 (10.9)	100 (18.2)	50 (9.1)
C	66	3	60 (10.9)	150 (27.3)	0 (0)
D	276	9 (+2)	240 (10.4)	400 (17.4)	100 (4.3)
E	259	4	60 (2.5)	330 (13.8)	0 (0)
F	90	0	0 (0)	0 (0)	0 (0)
G	162	14	250 (17.4)	700 (48.6)	0 (0)
H	120	6	100 (8.2)	200 (16.4)	50 (4.1)
I	120	3 (+1)	60 (4.9)	200 (16.4)	0 (0)
J	78	3	60 (8.8)	160 (23.5)	0 (0)
Total	1303	49 (+3)	970 (8.3)	2390 (20.4)	250 (2.1)

[a] A, rejected due to accuracy criterion; P, rejected due to precision criterion.

tors. A comparison between these approaches is shown in Table III for each of the analytes described earlier. If a criterion of ≤15% RSD were set for acceptance of the data, all methods would be considered acceptable except for analytes E and G at the level of the lowest control for the between-run data as analyzed by ANOVA. A ≤20% criterion for the lowest control as recommended by the joint conference report (1) would result in acceptance of all data regardless of how they were analyzed. Confounding of the data serves as a comparable assessment if the recommended criteria are applied, although information regarding the sources of imprecision are lost. Acceptability of the data depends on the limit set for %RSD and it appears that an acceptable procedure would be to evaluate precision with analysis of between- and within-run data separately. Since analytical runs are generally established with multiple control samples per run, it may be appropriate to analyze these data by ANOVA as long as it is realized that the two approaches may not be equivalent and that rejection criteria should reflect this.

Data Acceptance Criteria

Several issues must be considered when establishing a control program that allows acceptance of data with a defined limit on quality. Among these are the scientific validity of the approach, the throughput efficiency, and the ability to detect errors. There is an obvious trade-off between efficiency and ability to detect errors since the probability of false rejection of data is inversely related to the probability of error detection (3). This depends on the specific criterion set as well as the procedure used. The confidence interval approach for acceptance of data is well established in other disciplines and fixed ranges have been shown to be either too liberal or too conservative, depending on the analytical method (2).

The manner in which acceptance criteria are applied should also be considered. If one regards an analytical run as a unit and accepts or rejects entire runs, it is possible that entire runs may be rejected when only part of the run is invalid. In addition, a criterion applied to an entire run may allow errors to occur that go undetected if the acceptance criterion is liberal enough. It is therefore more scientifically valid to apply a criterion to each quality control which has been equally spaced throughout a run as in the bracketed approach and reject all samples on either side of that control rather than to allow a certain number of controls per run (one-third is the recommended and generally accepted portion) (1) to be outside acceptable ranges. A common criticism of the bracketed approach is that throughput efficiency may be sacrificed.

In order to address these issues, we have applied a 99% confidence interval with and without a separate accuracy criterion, a ±20% fixed range approach, a multiple rules confidence interval approach, and a range chart for individuals approach to both a Brackets and a Two-Thirds on Run approach for acceptance criteria. The results of this comparison for the analytes of interest are shown in Tables IVA–E.

The 99% confidence interval approach (Table IVA) resulted in 14 rejected controls from all runs. This resulted in rejection of 3.7% of samples by the bracketed approach and 2.9% by the Two-Thirds of Run approach. The data indicate that even when only precision is considered, more samples are rejected by the bracketed approach. If the accuracy criterion is applied sequentially before the precision criterion (Table IVB), the number of controls rejected almost triples and the number of rejected samples by the Bracketed approach almost doubles. The same criterion applied to Two-Thirds of Run, however, actually results in a decrease in the number of sample results that were judged unacceptable. This indicates that the Two-Thirds of Run approach is insensitive to detection of short-term errors. The Two-Thirds of Run approach results in fewer samples that require repeats, however, and is therefore more throughput efficient.

The fixed range criterion (Table IVC) is an accuracy criterion alone and essentially does not detect imprecision except as an additive effect on inaccuracy. This is a very simple approach which has advantages as mentioned previously but suffers in terms of the amount of information provided and the scientific validity of the approach. Table IVC shows that, again, more sample results are judged unacceptable by the Bracketed approach. Compared to Table IVB, however, it becomes evident that the fixed range approach provides a more stringent criterion than the accuracy/confidence interval approach. The difference in both the number of controls rejected and the samples rejected is not as dramatic as between the Bracketed and the Two-Thirds of Run approach. This indicates that the price in throughput efficiency that is sacrificed for the Bracketed approach may be compensated for when combined with the confidence interval approach.

The multiple rules approach proposed by Westgard *et al.* (9) is evaluated in Table IVD. Again, the bracketed approach is much more conservative than the Two-Thirds of Run approach, and by this criterion the multiple rules approach appears to be less stringent than either the accuracy/confidence interval or the fixed range approach. The Westgard approach was developed for application to an entire run of data, however, and works best for large runs. The number of samples repeated with the entire run judged by the multiple rules criterion is roughly comparable to the accuracy/confidence interval approach when run brackets are used for acceptance. The Westgard approach has been widely justified as optimal in terms of probabilities of error detection and false rejection (5–9).

The final approach evaluated is based on control charts for individuals (10). This approach establishes criteria based on both average and individual ranges. Average ranges are used as a criterion for run controls, whereas the individual range criterion is basically a repeatability assessment of duplicate controls. The criteria are applied sequentially. This approach has the advantage of statistical validity and maintains the capability to be applied to a small number of runs. The number of controls rejected and samples repeated when run brackets are considered is again comparable to the accuracy/confidence interval approach (Table IVE). This criterion applied to Two-Thirds of Run, however, results in the most liberal criterion except for Westgard's approach when applied in the same way.

CONCLUSIONS

Validation of accuracy and precision can be effectively evaluated by criteria established on %DFA and %RSD data, respectively. While the *t*-test approach proved to be a more stringent test for accuracy, the question is not whether there is any statistical difference but whether there is an acceptable difference from actual concentrations. Precision of an analytical method can be effectively evaluated between and within runs using unconfounded data with an acceptance limit of $\leq 15\%$. Within- and between-run data are often collected together in a routine situation but should not be confounded for precision assessment because information would be lost. Also, method validation should be considered apart from the routine analysis of samples. Sources of variation can be determined for confounded data by use of ANOVA, although results may not be equivalent to unconfounded data and a different criterion should be established.

Of the methods evaluated for data acceptance criteria, the Two-Thirds of Run criterion has been justified scientifically only for a fixed range rejection criterion of ±10% (4). It is, therefore, scientifically inappropriate to apply this approach to a ±20% fixed range as suggested by the joint conference report (1). The Two-Thirds of Run approach cannot be statistically justified for the confidence interval, multiple rules, or range approaches as well. For the three most statistically valid approaches (the multiple rules approach applied to the entire run and the accuracy/confidence interval and individual range approaches applied to run brackets), the number of samples repeated is comparable. The multiple rules approach requires at least 10 runs, the range approach requires 4, and the confidence interval approach can be used with any number of runs, with a larger number providing better estimates of mean values and acceptance range. The confidence interval approach is simpler and easier to apply than multiple rules or individual range approaches, although it would seem that similar results can be obtained with appropriately applied individual range or multiple rules approaches. The ±20% fixed range approach applied to Two-Thirds of Run recommended by the joint conference (1) is difficult to justify scientifically and appears to be a significantly more liberal criterion than the other procedures evaluated.

The additional precision and accuracy criterion recommended by the joint conference was interpreted to be a criterion applied to an entire set of runs and was not evaluated as a run acceptance criterion. The fifteen and twenty percent criterion for accuracy and precision should serve as a good monitor of method performance as data are accumulated, although it should be understood that application of a criterion to multiple runs is largely a validation and not a quality control issue.

REFERENCES

1. V. P. Shah, K. K. Midha, S. Dighe, I. J. McGilveray, J. P. Skelly, A. Yacobi, T. Layloff, C. T. Viswanathan, C. E. Cook, R. D. McDowell, K. A. Pittman, and S. Spector. Analytical methods validation: Bioavailability, bioequivalence and pharmacokinetic studies. *Pharm. Res.* 9:588–592 (1992).
2. H. T. Karnes, G. Shiu, and V. P. Shah. Validation of bioanalytical methods. *Pharm. Res.* 8:421–426 (1991).
3. J. O. Westgard and P. L. Barry. *Cost Effective Quality Control: Managing the Quality and Productivity of Analytical Processes*, AACC Press, Washington, DC, 1986, pp. 33–64.
4. A. G. Causey, H. M. Hills, and L. J. Phillips. Evaluation of criteria for the acceptance of bioanalytical data. *J. Pharm. Biomed. Anal.* 8:625–628 (1990).
5. C. A. Parvin. Comparing the power of quality-control rules to detect persistent increases in random error. *Clin. Chem.* 38:364–369 (1992).
6. C. A. Parvin. Comparing the power of quality-control rules to detect persistent increases in random error. *Clin. Chem.* 38:358–363 (1992).
7. J. O. Westgard and T. Groth. A predictive value model for quality control: Effects of the prevalence of errors on the performance on control procedures. *Am. J. Clin. Pathol.* 80:49–56 (1983).
8. J. O. Westgard, T. Groth, and C.-H. de Verdier. Principles for developing improved quality control procedures. *Scand. J. Clin. Lab. Invest.* 44(Suppl. 172):19–41 (1984).
9. J. O. Westgard, P. L. Barry, M. R. Hunt, T. Groth, R. W. Burnette, A. Harnline, R. E. Thiers, and H. Nipper. A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin. Chem.* 27:493–501 (1981).
10. J. R. Lang and S. Bolton. A comprehensive method validation strategy for bioanalytical applications in the pharmaceutical industry. 2. Statistical analyses. *J. Pharm. Biomed. Anal.* 9:435–442 (1991).
11. H. T. Karnes and C. March. Calibration and validation of linearity in chromatographic biopharmaceutical analysis. *J. Pharm. Biomed. Anal.* 9:911–918 (1991).