BioPharm INTERNATIONAL
The Science & Business of BioPharmaceuticals

Analytical Methods: A Statistical Perspective on the ICH Q2A and Q2B Guidelines for Validation of Analytical Methods

Dec 1, 2006

By: Steven Walfish
BioPharm International

ABSTRACT

Vagueness in the ICH Q2A and Q2B guidelines necessitates effective protocol design and data analysis. For specificity (detection in the presence of interfering substances), the goal is statistical differences with meaningful implications on assay performance. Linearity (results directly proportional to concentration of analyte in the sample) is typically demonstrated via least squares regression. Accuracy (difference between measured and true values) usually is presented as a percent of nominal. Precision analysis is vital because it supports claims of accuracy and linearity. A well-designed experiment and statistically relevant methods will facilitate method validation in accordance with ICH guidelines.

Several articles have been published on the requirements of method validation for analytical methods.[1,2] Most of these articles do not, however, concentrate on the protocol design and analysis of data from these studies. The International Conference on Harmonization (ICH) guidelines on Validation of Analytical Procedures (Q2A and Q2B) delineate the guidance and methodology for validation characteristics of an analytical procedure, but as in many guidelines, the terminology is vague enough to allow for several acceptable approaches and analyses. Appropriate statistical methods should be used; in addition, all relevant data collected during validation and all formulae used for calculating validation characteristics should be submitted and discussed as appropriate.

The following excerpt from the ICH Q2B guideline is an example of the vagueness that can trouble many scientists:

Approaches other than those set forth in this guideline may be applicable and acceptable. It is the responsibility of the applicant to choose the validation procedure and protocol most suitable for their product. However, it is important to remember that the main objective of validating an analytical procedure is to demonstrate that the procedure is suitable for its intended purpose. Due to their complex nature, analytical procedures for biological and biotechnological products in some cases may be approached differently than in this document.[3]

The ICH guidelines suggest combining individual validation characteristics to minimize total testing. A statistical approach to validation of analytical methods can minimize the amount of testing while meeting the requirements of the guidelines. This assertion is based on the following comment from the ICH Q2B document:

In practice, it is usually possible to design the experimental work such that the appropriate validation characteristics can be considered simultaneously to provide a sound, overall knowledge of the capabilities of the analytical procedure, for instance: specificity, linearity, range, accuracy and precision.[3]

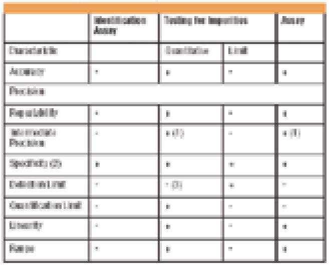
TYPES OF ANALYTICAL METHODS

There are four common types of analytical methods, each with its own set of validation requirements. The level of stringency is proportional to the criticality of the method in testing drug product. The four most common types of analytical procedures are:

- Identification tests
- Quantitative tests for impurities' content
- Limit tests for the control of impurities
- Quantitative tests of the active moiety in samples of drug substance, drug product, or other selected component(s) in the drug product.

The elements of the analytical method requiring proof through validation as contained in the ICH Q2A guideline are summarized here in Table 1.[4]

Table 1. Assuming one needed to show acceptable results for all characteristics, a testing scheme could be developed to meet the minimum test requirements of the guidance.



	Identification Assay	Testing for Impurities		Assay
Characteristic		Quantitative	Limit	
Accuracy	-	+	-	+
Precision				
Repeatability	-	+	-	+
Intermediate Precision	-	+ (1)	-	+ (1)
Specificity (2)	+	+	+	+
Detection Limit	-	- (3)	+	-
Quantitation Limit	-	+	-	-
Linearity	-	+	-	+
Range	-	+	-	+

Table 1. Assuming one needed to show acceptable results for all characteristics, a testing scheme could be developed to meet the minimum test requirements of the guidance.

SPECIFICITY

Specificity usually is defined as the ability to detect the analyte of interest in the presence of interfering substances. Specificity can be shown by spiking known levels of impurities or degradants into a sample with a known amount of the analyte of interest. A typical testing scheme would be to test a neat sample and a minimum of three different levels of interfering substances. Several different analysis methods have been proposed to determine specificity; these include percent recovery, minimum difference from baseline, and analysis of variance. Currently, there are differences in opinion regarding the appropriateness of using analysis of variance (ANOVA) for showing a difference between baseline and a spiked sample. The goal is not to find statistically significant differences that have no practical value, but to find statistical differences that have meaningful implications on assay performance. It is common in clinical diagnostics to use a *t*-test to assess sensitivity (minimum detected dose or concentration), specifically using a method by Rodbard.[5]

One proposed method, which combines both the statistical rigor of analysis of variance and the appropriateness of meaningful differences from baseline, is to use equivocal tests or a method similar to the one used to assess parallelism.[6] In this method, the comparison must be within the equivocal zone, though not statistically different. Figure 1 shows four scenarios; in each of these, the equivocal zone is determined by the distance between –λ and +λ, which is the predetermined level that is scientifically not different than the target.

1. In scenario 1, the 95% confidence interval (denoted by the horizontal line) contains the target, and the entire 95% confidence interval is contained in the equivocal zone. In this case, both statistical significance and scientific judgment agree.

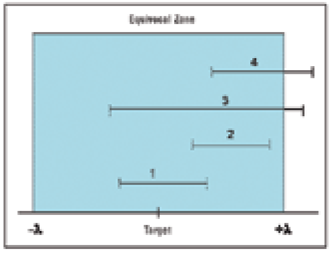


Figure 1. Four scenarios, using the equivocal method. In each, the equivocal zone is determined by the distance between –λ and +λ, which is the predetermined level that is scientifically not different from the target.

Figure 1

2. In scenario 2, the 95% confidence interval does not contain the target; therefore, it would be considered statistically different, although the 95% confidence interval is fully contained in the equivocal zone. In this case, one would judge the sample to be scientifically similar to the target.

3. In scenario 3, the 95% confidence interval would lead one to conclude there is no statistical significance, but the 95% confidence interval is not fully contained in the equivocal zone. Because the variability is larger here, one cannot conclude there is a statistical difference, but scientifically it is shown to be possibly too large a difference.

4. In scenario 4, neither the 95% confidence interval nor the equivocal zone shows that the sample is equivalent to the target.

In scenarios 1 and 4, both methods agree, whereas scenarios 2 and 3 have some discrepancy. In scenario 2, the precision is so good that the statistical test fails, although in a practical sense it is similar to the target. Scenario 3 gives the most confusing conclusion; because the confidence interval is not fully contained in the equivocal zone, one might increase sampling or perform a retest to attempt to reduce the variability. There is no clear answer for this scenario. Unfortunately, the selection of the equivocal zone and associated λ value is still being debated in the statistical and scientific community. A potential compromise is to use some percentage, such as 75% of the specification width, as the equivocal zone for specificity testing.

A minimum of three repeat readings should be taken for each sample; the ideal would be six repeats. Increased repeat readings of a sample give the analysis of variance more power to detect a difference, if a difference exists.

LINEARITY AND RANGE

The linearity of an analytical procedure is its ability, within a given range, to obtain test results that are directly proportional to the concentration of analyte in the sample. De facto, the range is the smallest and largest concentration that maintains a linear relationship between the concentration and the response of the method. The ICH guidelines do not require any proof of precision, though it is clear that without sufficient precision, the linear relationship cannot be guaranteed. The most common method used for demonstrating linearity is least squares regression. Sometimes it is necessary to transform the data to get a linear fit. The guidelines recommend a minimum of five dose levels throughout the range, with each tested for a minimum of three independent readings. It is possible to use these samples to test the accuracy of the method. Accuracy is the lack of bias at each level, and as long as the bias is consistent along the range of the assay, the method is considered linear. Residual analysis, or the observed value minus the predicted (from the linear equation) can help to assess if there is sufficient linearity in the data. Daniel and Wood give an excellent explanation of residual analysis.[7]

ACCURACY

Accuracy is the difference between the measured value and the true value. This is different from trueness, which is the difference between the mean of a set of measured values and the true mean value. Accuracy is usually presented as a percent of nominal, although absolute bias is also acceptable. Accuracy in the absence of precision has little meaning. Accuracy claims should be made with acceptable precision. ICH guidelines suggest testing three replicates at a minimum of three concentrations. If the same data from the linearity experiment are used, then there would be five levels. Precision of the data should be compared to observed precision from previous studies or development runs, to confirm the observed precision and validity of the accuracy runs.

ICH guidelines recommend using confidence intervals for reporting accuracy results. Confidence intervals are used for probability statements about the population mean—for example, that the average percentage recovery should be 95–105%.

Tolerance intervals can be used to set appropriate accuracy specifications. These say, for example, that no individual percentage recovery should be less than 80% or greater than 120%.

Tolerance intervals make a statement about the proportion of the population values with a fixed confidence. Therefore, one would say that *x*% of the population will be contained in the tolerance limits with *y*% confidence. Tolerance intervals are computed from the sample mean and sample standard deviation. A constant *k* is used such that the interval will cover *p* percent of the population with a certain confidence level. The general formula for a tolerance interval is: *x-mean* ± *kS*

Values of the *k* factor as a function of *p* and percent confidence are tabulated in Dixon and Massey.[8]

PRECISION

The most important part of any analytical method validation is precision analysis. The ICH guidelines break precision into two parts: repeatability and intermediate precision. Repeatability expresses the precision under the same operating conditions over a short interval of time. Repeatability is also termed *intra-assay precision*. Intermediate precision expresses within-laboratory variations: different days, different analysts, different equipment, etc. Additionally, the ICH Q2A guideline defines reproducibility as the precision among laboratories (collaborative studies, usually applied to standardization of methodology).[4] This lab-to-lab precision could be combined into the estimate of intermediate precision, because it is possible that a particular test method could be run in more than one laboratory. The suggested testing consists of a minimum of two analysts on two different days with three replicates at a minimum of three concentrations. If lab-to-lab variability is to be estimated, the experimental design should be performed in each lab. The analyst and day variability combine to give the intermediate precision (lab-to-lab, if estimated, is added here), whereas the variation after accounting for the analyst and day is the repeatability.

Variance components, or decomposition of variance, is a statistical method to partition the different sources of variation into their respective components. Statistical programs such as Minitab are commonly used to calculate variance components. In Minitab, the option to calculate variance components is contained in the analysis of variance (ANOVA) menu option. Box, Hunter, and Hunter provide an excellent source for additional information on how to calculate variance components.[9] It is important to remember that variances can be added or averaged, but not the standard deviations.

DETECTION AND QUANTIFICATION LIMITS

The detection limit of an assay is the lowest concentration that can be detected but necessarily quantified; the quantification limit is the lowest concentration that can be quantified with acceptable precision. The quantification limit is the lowest level of analyte that can be reported. The ICH guidelines suggest three different methods for determining the detection and quantification limits. These are: visual determination, signal-to-noise determination, and standard deviation and slope method. Each method will give different results. The signal-to-noise method is the most logical, because it is based on comparing low levels of the analyte to a blank or background sample.

Determination of the signal-to-noise ratio is performed by comparing measured signals from samples with known low concentrations of analyte with those of blank samples, and subsequently establishing the minimum concentration at which the analyte can be reliably detected. A signal-to-noise ratio between 3:1 and 2:1 is generally considered acceptable for estimating the detection limit.[3]

SUGGESTED PROTOCOL

Using a well-designed experiment can reduce the total testing time. A well-designed experiment can also improve the quality of the analysis by improving the statistical power. Certain characteristics need to be tested individually because the sample preparation is unique for the test. Assuming one needed to show acceptable results for all characteristics in Table 1, a testing scheme could be developed to meet the minimum requirements of the guidance. Three experiments would cover all the characteristics and their minimum sample requirements.

The first experiment is for specificity alone. A minimum of three different levels for each potential interfering substance, plus a neat sample, is repeated six times for a total of 24 repeats for the experiment. A one-way analysis of variance is used to test the level that can be shown to be statistically different from the neat sample. The last level of the interfering substance that is shown to be statistically equivalent to the neat sample is the level of specificity. Specificity testing should be conducted for all potential interfering substances. It is not necessary to spike levels of interfering substances that cannot reasonably be expected to be present in test samples.

The second experiment establishes the detection limit and quantification limit. This experiment is run only for those assays that require such characterization. If one does not expect the range of the assay to test samples near the quantification level, this experiment can be eliminated. As stated above, there are several ways to demonstrate detection and quantification limits. Here we propose a test scheme using the signal-to-noise method of a blank sample. Testing eight repeats of the blank sample gives a sufficient estimate of the error. Using the standard deviation of the blank sample, the detection and quantification limits are set. The mean of the eight repeats, plus three times the standard deviation, is the detection limit, while 10 times the standard deviation is the quantification limit. If additional precision around these estimates is necessary, the number of repeats is increased. The signal-to-noise method can be used to assess the detection limit, and one of the other methods can be used to assess quantification limit. Regardless of the method used for assessing quantification limit, a sample at that level should be incorporated into experiment three.

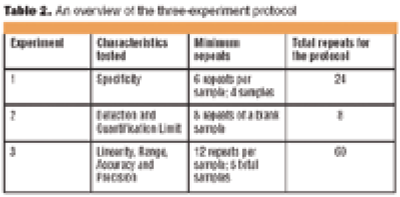


Table 2. An overview of the three-experiment protocol

Experiment	Characteristics tested	Minimum repeats	Total repeats for the protocol
1	Specificity	6 repeats per sample; 4 samples	24
2	Detection and Quantification Limit	8 repeats of a blank sample	8
3	Linearity, Range, Accuracy and Precision	12 repeats per sample; 5 total samples	60

Table 2. An overview of the three-experiment protocol

The third experiment covers the majority of characteristics required by the guidelines. After the data from experiment three is analyzed, this data should be compared to the levels of interfering substances and detection limit to show acceptable precision of the test method at these levels. A minimum of five samples spanning the expecting range of the assay are each tested a minimum of 12 times (two analysts over two days testing three replicates per day). Usually, multiple analysts and days are used to estimate intermediate precision. A minimum of two analysts will perform the assay on a minimum of two days, with three repeats on each day (a total of 12 observations per sample). To meet the linearity requirement, a minimum of five samples will be used. Table 2 gives an overview of the three-experiment protocol discussed here. More samples and repeats can be added to the study to gain additional information about the test method.

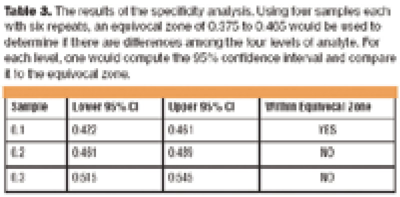


Table 3. The results of the specificity analysis. Using four samples each with six repeats, an equivocal zone of 0.375 to 0.465 would be used to determine if there are differences among the four levels of analyte. For each level, one would compute the 95% confidence interval and compare it to the equivocal zone.

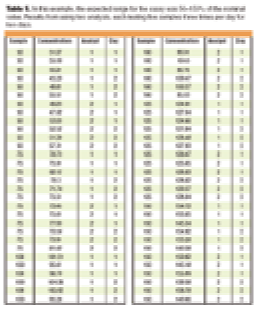
Sample	Lower 95% CI	Upper 95% CI	Within Equivocal Zone
0.1	0.422	0.451	YES
0.2	0.451	0.488	NO
0.3	0.515	0.545	NO

Table 3. The results of the specificity analysis. Using four samples each with six repeats, an equivocal zone of 0.375 to 0.465 would be used to determine if there are differences among the four levels of analyte. For each level, one would compute the 95% confidence interval and compare it to the equivocal zone.

The use of standards, product, and spiked excipients for method validation depends on the purpose of the method under test. If National Institute of Standards and Technology (NIST) or World Health Organization (WHO) standards exist for the test method, or if a pharmacopeial method is under consideration, those standards should be used. Standards can be used if the test method requires a standard curve to calculate the response. Bulk drug product or final drug product are used when stability indicating assays are being tested. Stressed samples and time zero samples typically are used for method validation of these stability-indicating assays. When novel proteins or unique products are being tested, typically the only samples that are available are spiked samples. When using spiked samples, one must assume that the spiked value is known without error.

Table 4. Based on the results of the specificity analysis (Table 3), we conclude that the specificity of the assay is 0.2 mg/dL.

EXAMPLE



The following example shows how specificity analysis would be conducted using the above protocol. Using four samples each with six repeats, an equivocal zone of 0.375 to 0.465 would be used to determine if there are differences among the four levels of analyte. For each level, one would compute the 95% confidence interval and compare it to the equivocal zone. Table 3 shows the results. Based upon those results, we conclude that the specificity of the assay is 0.2 mg/dL (Table 4).

Table 5. In this example, the expected range for the assay was 50–150% of the nominal value. Results from using two analysts, each testing five samples three times per day for two days.

The validation of linearity, accuracy precision, and range was performed on a cell-based bioassay that did not require a level of quantification or level of detection. The expected range for the assay was 50–150% of the nominal value. Using two analysts, each testing five samples three times per day for two days yielded the results in Table 5.

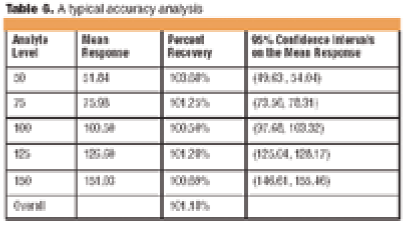


Table 6. A typical accuracy analysis

Analyte Level	Mean Response	Percent Recovery	95% Confidence Intervals on the Mean Response
50	51.84	103.68%	(49.63, 54.04)
75	75.93	101.23%	(73.56, 78.31)
100	100.50	100.50%	(97.68, 103.32)
125	126.58	101.28%	(125.04, 128.17)
150	151.03	100.69%	(146.61, 155.46)
Overall		101.18%	

Table 6. A typical accuracy analysis

Once the data are collected, one has the proverbial chicken-or-egg decision; in this case, accuracy or precision. Since accuracy and precision go hand-in-hand, the decision of which to assess first involves personal preference. Here, for the sake of simplicity, accuracy will be assessed first, then precision. Accuracy is calculated by combining all data across analysts and days for each level of analyte. Depending on the method being validated, acceptance criteria should be established. For the following example, one can use a percent recovery of 95–105%. A typical accuracy analysis is shown in Table 6.

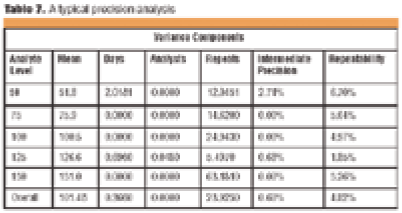


Table 7. A typical precision analysis

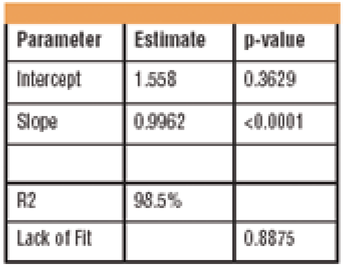
		Variance Components				
Analyte Level	Mean	Days	Analysts	Repeats	Intermediate Precision	Repeatability
50	51.8	2.0181	0.0000	12.3451	2.78%	6.39%
75	75.3	0.0000	0.0000	14.6290	0.00%	5.04%
100	100.5	0.0000	0.0000	24.3430	0.00%	4.57%
125	126.6	0.0960	0.0000	5.4370	0.63%	1.85%
150	151.0	0.0000	0.0000	65.1840	0.00%	5.36%
Overall	101.45	0.7600	0.0000	25.8250	0.67%	4.85%

Table 7. A typical precision analysis

Without a precision analysis, one cannot confirm accuracy claims. The intermediate precision includes analyst and day, while the repeatability includes the variability within analyst per day. Each source of variability is assessed and then combined to yield the intermediate precision and repeatability. A typical precision analysis is contained in Table 7.

Once one has shown acceptable precision and accuracy, one can assess if the bias is constant by performing a linearity analysis. Using the same data from the accuracy and precision analysis, an ordinary least squares (OLS) estimate can be calculated. Two coefficients are estimated using OLS: the slope and intercept. A lack-of-fit test confirms that the linear model is appropriate for the data set. Combining all the data similar to the accuracy analysis yields the linearity analysis contained in Table 8. A graphical representation of the data is shown in Figure 2. The model illustrates a statistically significant slope with a lack-of-fit test showing that the linear model is sufficient (for lack-of-fit test, a *p*-value greater than 0.05 is indicative that the model is sufficient). The intercept is not statistically significant ($p > 0.05$), indicating that the assay would run through the origin.

Table 8. A typical linearity analysis



Parameter	Estimate	p-value
Intercept	1.558	0.3629
Slope	0.9962	<0.0001
R2	98.5%	
Lack of Fit		0.8875

Table 8. A typical linearity analysis

Since the accuracy, precision, and linearity all meet the requirements, one can state that the range of the assay is 50–150%.

SUMMARY

Using a well-designed experiment and statistically relevant methods, method validation can be accomplished in accordance with the ICH guidelines. Precision analysis is the most critical component because it allows the claims of accuracy and linearity to be made.

Figure 2. Graphical representation of data from the typical linearity analysis, Table 8.

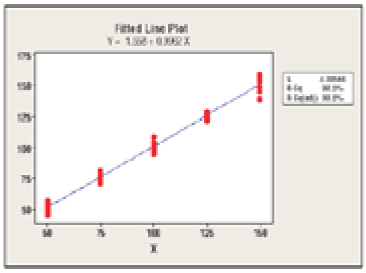


Figure 2

Steven Walfish is president of Statistical Outsourcing Services and *BioPharm* Editorial Board Member, 403 King Farm Boulevard, Suite 201, Rockville, MD 20850, 301.325.3129, fax: 301.330.2143, steven@statisticaloutsourcingservices.com

REFERENCES

1. Krause SO. Analytical method validation for biopharmaceuticals: a practical guide. Guide to Validation. Supp to BioPharm Int. 2005 Mar; 26–34.

2. Kanarek AD. Method validation guidelines. Guide to Bioanalytical Advances. Supp to BioPharm Int. 2005 Sep; 28–33.

3. US Food and Drug Administration. Guidance for industry: Q2B validation of analytical procedures: methodology. Rockville, MD: Nov 1996.

4. US FDA. Guideline for industry: text on validation of analytical procedures: ICH Q2A. Rockville, MD: Mar 1995.

5. Rodbard D. Statistical estimation of the minimal detectable concentration (sensitivity) for radioligand assays. Analytic Biochem. 1978;90:1–12.

6. Hauck WW, Capen RC, Callahan JD, De Muth JE, Hsu H, Lansky D, et al. Assessing parallelism prior to determining relative potency. PDA J Pharma Sci Technol. 2005;59(2):127–37.

7. Daniel C, Wood FS. Fitting equations to data. 2nd ed. New York: Wiley & Sons; 1980.

8. Dixon WJ, Massey FJ. Introduction to statistical analysis. New York: McGraw-Hill; 1969.

9. Box GEP, Hunter WG, Hunter JS. Statistics for experimenters. New York: Wiley & Sons; 1978.