ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
28 (2002) 251–260

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

# Validation of a near-infrared transmission spectroscopic procedure, part A: validation protocols

Howard Mark [a,*], Gary E. Ritchie [b], Robert W. Roller [c], Emil W. Ciurczak [b], Cindy Tso [b], Stacy A. MacDonald [b]

[a] *Mark Electronics, 21 Terrace Avenue, Suffern, NY 10901, USA*
[b] *Purdue Pharma LP, Building 444, Saw Mill River Road, Ardsley, NY 10502, USA*
[c] *Eli Lilly and Company, Indianapolis, IN 46285, USA*

Received 7 May 2001; received in revised form 8 August 2001; accepted 26 September 2001

**Abstract**

It is possible to devise calibration and validation protocols that enable the ICH guidelines to conform to the specialized requirements of the NIR method of analysis. Some of the required characteristics of evaluation specified by the guidelines, such as accuracy and repeatability, can be applied directly, just as with any other analytical method. Other characteristics are adapted through the novel use of specialized statistics, or through the use of creative methods and procedures to match the recommendations of the guidelines to the unique and specialized requirements of the NIR method. 

*Keywords:* Validation; Near-infrared transmission spectroscopy; Release assay; Chemometrics

## 1. Introduction

As an analytical tool for the pharmaceutical analytical laboratory, near-infrared spectroscopic methods have the potential to dramatically improve the quality of the drug manufacturing process from assessing incoming raw materials to the final drug product. Currently, other spectroscopic methods, chromatographic, titrimetric and wet methods have been and are being used to analyze bulk drug and finished products. These methods are labor intensive, costly and typically require two or more working analyst days to complete. In addition, organic solvent waste is often generated by these methods, creating expense, disposal problems and potential safety hazards. NIR methods in contrast, can reduce the costs of testing, require no reagents, associated reagent preparation steps, sample preparation steps, and generally require only one working analyst day to complete testing. Since no reagents are required, no additional costs are incurred for solvent waste disposal and there are few, if any, hazards associated with NIR techniques. The instrumentation used is eminently suited to being used in production facilities, and on-line measurements are routinely used in various industries [1,2].

* Corresponding author. Tel.: +1-845-357-9496; fax: +1-845-369-9110.
*E-mail address:* hlmark@prodigy.net (H. Mark).

0731-7085/02/$ - see front matter 
PII: S0731-7085(01)00567-2

Due to these beneficial characteristics, NIR analysis has long enjoyed widespread use in the agricultural [3] and other industries. These industries have the common characteristic of being non-regulated. When introducing NIR analysis into a regulated industry, such as the pharmaceutical industry, special attention must be paid to satisfying the regulatory requirements for analytical procedures. The US Pharmacopoeia has proposed guidelines [4], which are currently in revision [5]. While no regulations have yet been published in the US, the European Pharmacopoeia has published a monograph [6] describing procedures for qualitative analysis applicable to transmitance measurements of non-scattering solutions and to diffuse reflectance measurements.

The guidelines of the International Conference on Harmonization (ICH), and the US Pharmacopoeia are generally used as the basis for developing methods that promote consistency, uniformity of regulatory effort and conformity to stringent standards for validation of analytical methods in the US. The FDA has issued guidelines based on the ICH recommendations [7,8] although these guidelines do not specifically address methods of NIR analysis.

The ICH guidelines contain definitions for the following characteristics of an analytical procedure [7]:

1. specificity;
2. linearity;
3. range;
4. accuracy;
5. precision;
6. repeatability;
7. intermediate precision;
8. reproducibility;
9. detection limit;
10. robustness.

Moffat et al. have recently reviewed the requirements for adhering to the standards that must be met in order to gain approval of NIR methods for use in the pharmaceutical industry [9]. As Moffat et al. point out, the guidelines from the US Pharmacopoeia and those of the ICH are largely complementary. The Pharmacopoeia guidelines concern themselves largely with those issues related to ensuring proper operation of the hardware and software before and during the measurement process. These procedures insure that the data collected is accurate and reliable. The pharmacopoeia guidelines constitute a codification of standard good NIR practice, and need no further attention here.

The ICH guidelines, on the other hand, tend to concentrate their attention on those matters that become of concern after proper operation of the hardware has been verified, that is, those issues dealing with the precision and accuracy of the final analysis. In other words, once there is an assurance that the data collected will be accurate and reliable, attention then turns to the question of what data to collect. Moffat, et al. also discuss the situations where not all of the above-listed characteristics apply to all types of analyses.

NIR analysis differs from other analytical methods in that it relies heavily on the use of chemometric calibration and statistical analysis of data to produce the benefits described above. For this reason, verifying the hardware operation, while important, does not suffice to insure long-term accuracy of the analytical results. Therefore, additional statistical tests are normally applied to the results from NIR data in order to provide implicit quality-control procedures. The American Society for Testing and Materials (ASTM) has created recommended practices that can be used as the basis for such testing; the one most pertinent here is ASTM practice # E1655 [10]. Due to the nature of the technology, these statistical testing methods must be used to compute the values of the various characteristics that the ICH guidelines specify.

Moffat et al. provide one example of how the guidelines may be met; however, they also take pains to point out that their procedures provide only one of many possibilities, and that other procedures may be equally suitable or even superior. Moffat's example was applied to data resulting from measurements of diffuse reflectance. While a different measurement technique does not automatically negate the possibility of applying Moffat's procedures, the interest in the current work was in the application of NIR analysis using diffuse transmittance; this application is described in a companion paper [11].

This paper is intended to present a methodology for assessing the suitability of an NIR method using diffuse transmittance as the measurement technique. When reading the original guidelines [7,8] it becomes clear that multivariate methods of analysis, such as NIR, were not considered in the development of the guidelines. None of the unique characteristics of such methods were taken into account, neither to take advantage of their unique benefits, nor to minimize any potential liabilities. However, the guidelines are general and flexible enough that they can conform to the requirements and limitations of the method. We hope that the methodology presented here can serve as a basis for generating standardized procedures for validation of NIR pharmaceutical analyses. At this time we, along with Moffat et al. [9], disclaim any intention that the methodology presented here is the only suitable one, nor that it should be followed slavishly, especially in variant situations.

## 2. Validation protocols

The Pharmacopoeia guidelines for validation of analytical methods include a table specifying the characteristics of a method that must be evaluated for analytical methods to be used for various purposes (see table 2 in [12]). The ICH guidelines contain an abbreviated version of this table laying out their recommendations as to which analytical characteristics are important for the various applications of analytical methods [7]. This table indicates that only those characteristics of the analytical method appropriate to the intended application need be verified. Our intended application for the use of these guidelines is for content uniformity and release assay analysis [11], both of which require the analysis of a major component in the samples. For major components, the guidelines recommend that the following characteristics are to be evaluated:

accuracy;
repeatability;
intermediate precision (alternatively, reproducibility);
specificity;
linearity;
range;
robustness.

As Moffat's excellent discussion describes, the remaining characteristics are not evaluated for major component analysis. This is because either they are not recommended by the ICH guidelines [9] or because they require an inter-laboratory study and the guidelines themselves proclaim those characteristics are not appropriate for market authorization approval. These are:

detection limit;
quantitation limit;
reproducibility.

In addition to the above characteristics, the guidelines also indicate that quantitative analytical methods should be accompanied by qualification procedures. Due to the known propensity for the NIR modeling procedure to sometimes 'key in' on characteristics of the calibration data other than the properties of analyte, it is important to ensure that it is in fact the analyte that is being used as the basis for the measurement. While the ICH Q2B guidelines do not define or use the term 'qualification procedure', it does contain the following passage:

> "The discrimination of a procedure may be confirmed by obtaining positive results (perhaps by comparison with a known reference material) from samples containing the analyte…".

We interpret this passage to mean 'a procedure that demonstrates that the instrument is responding to the actual analyte' and we use the term 'method qualification procedure' as a shorthand term for that. We believe this is an important procedure to ensure the reliability of the NIR measurement. While this passage is contained in the section of the ICH guidelines labeled Specificity we believe this is distinct from 'specificity', which we understand to mean a lack of response to materials other than the desired analyte.

Robustness should be considered during the development phase for an analytical method but is not commonly evaluated at that time. Therefore, it may or may not be built in during method

development. We now present the methodology used to evaluate the necessary characteristics in the current study:

### 2.1. Accuracy

The ICH guidelines recommend using a minimum of nine samples, three samples at each of three concentration levels, this is the same recommendation as for specificity. Three methods are provided for determining the accuracy from these data. Of the three, the one most applicable to NIR analysis is 'comparison of the results of the proposed analytical procedure with those of a second well-characterized procedure, the accuracy of which is stated and defined'. The guidelines, however, do not specify how to make the comparison or how to calculate the accuracy. The methodology of NIR analysis, however, inherently addresses these issues.

One of the tenets of NIR analysis is the requirement for multivariate calibration of the instrumental response against a set of calibration samples, whose concentrations are known from another analytical technique, a 'reference method'. Use of an approved analytical method as the reference method inherently satisfies the requirement that the accuracy of the reference method be known.

Typical practice for NIR analysis indicates that the quantity of calibration samples used is between twenty and several hundred. These samples are typically spread out over many more than three concentrations. These characteristics of the calibration process also constitutes an inherent parameter that well satisfies the ICH recommendation.

The calibration procedure invariably uses one of three widely used mathematical algorithms to relate the instrument readings to the concentration information. These algorithms are:

Multiple Linear Regression (MLR), also known as Inverse Beer's Law, Inverse Least Squares, and by several other names.

Principal Component Regression (PCR).

Partial Least Squares (PLS).

All three of these algorithms normally calculate several auxiliary statistics during the course of the computations. These calculations are specified by the ASTM practice # 1655-97 for quantitative NIR analysis [10]. One of these auxiliary statistics is the standard error of calibration, perhaps more familiar to some as the standard error of estimate (S.E.E.).

Another auxiliary statistic defined in ASTM practice # 1655 is the standard error of validation (S.E.V.), again more familiarly known as the standard error of prediction (S.E.P.). This is a calculation similar to that of the S.E.C., but applied to data from samples not included among the samples upon which the calibration calculations were performed. That same data can be used to evaluate the bias of the NIR method by computing the average difference between the NIR values and the values from the reference method for the same samples.

Another statistic that provides a valuable comparison between results from the NIR method and the results from the reference method is the bias of the readings from the independent set of validation samples. The bias is the average (arithmetic mean) of the differences between the two methods, and should be small. The bias calculation is not applicable to the data from the calibration samples, however.

Other calculations can be performed but we forego discussion of the details here. The above-described calculations (S.E.E. and S.E.P.) produce numbers that are effectively the standard deviation of the differences between the NIR and reference technique, and thus they provide a method of performing the comparison that is objective and statistically sound. Combined with the bias computed from the validation data, these give a reliable picture of the accuracy of the NIR method.

### 2.2. Repeatability

The ICH guidelines describe two alternative methods of ascertaining repeatability:

1. Minimum of six readings of a single sample at 100% of target concentration.
2. Minimum of three readings on each of three samples, one at each of three levels of concentration.

Of these, the one easier to apply to NIR measurements is the result of at least six measurements at 100% of the target concentration.

The procedure for calculating repeatability given in the ASTM practice [10] is basically the same as the ICH recommendation, except that the ASTM practice also specifies how to calculate the repeatability of the measurements of the sample.

As with most recommendations in the ICH guidelines, the meaning of repeatability is based on analytical procedures based on chemistry or chromatography. In the case of NIR analysis, there are possible variations that can occur, that would affect the way 'repeatability' is measured. For example, tablets can be removed from a holder and replaced in the same position, or not removed at all. Powders can be measured multiple times without being disturbed, or the same aliquot of powder can be similarly removed/replaced in a holder or sample cup. Any of these scenarios could legitimately be considered a 'repeatability' measurement. It is up to the scientist developing the method to decide what 'repeatability' means to him, which will thus determine how it should be measured.

To use our own work as an example of one way to perform this measurement, we measured three samples, at 80, 100 and 120%, of the target concentration. Each sample had its spectrum measured 13 times without moving the sample and the analytical value corresponding to each of these spectra calculated from the calibration model in use. The repeatability is then calculated as the standard deviation (S.D.) of these values [10].

Other combinations of number of samples, number of repeat readings per sample, etc. are equally valid ways to measure repeatability, as long as they meet the minimum ICH requirement listed above.

### 2.3. Intermediate precision

The ICH recommendation for intermediate precision is to study the effect of random events on the readings. While examples of typical variations are given, a list that includes different analysts and different days, the nature of the random events that need to be studied are not specified,

Again using our own work as an example of what can be done (although other ways of meeting the requirements of the guidelines are also possible, and other sources of variability must be considered when they are present), the experiments described in the companion article were carried out, and the instrumentation is intended to be used in a well-controlled laboratory environment. Thus other potential sources of random variation, such as variation of environmental temperature are not present. Therefore, in the study to which this protocol is applied, the effect of the random variables analyst and days were investigated.

For each of the random variables, data was collected using a protocol similar to that for repeatability. A sample at 100% of the target concentration was measured 10 times at the variant conditions, after having been similarly measured under the initial conditions for the repeatability measurement.

### 2.4. Range

The acceptance range for an assay is defined by the ICH [7]. The minimum range to be used in the validation of an analytical method is specified in [8] as 80–120% of the target concentration for assay and 70–130% of the test concentration for content uniformity testing. Moffat [9] discusses the situations where meeting those requirements is physically impossible and/or impracticable.

In addition to the limitations described by Moffat, there is a mismatch between the ICH guidelines and the technical requirements of NIR analysis. Good NIR practice dictates that samples to be used for calibrating the instrument should be the same as the samples that are to be analyzed in the future. In the case of samples that are the product of a defined production process, this requirement normally means that samples to be used for calibration, as well as, the samples used for testing the analytical method should come from that production process. The problem with this method of obtaining samples is that samples from a pharmaceutical production process in good control include only a very limited range of analyte concentration, usually a far smaller range

than the ICH guidelines specify. It is generally illegal to deliberately produce samples that cover the range specified for analysis. Occasional process upsets may produce samples covering a wider range. Such samples, however, are likely to differ from 'normal' production samples in other ways. The result of this is that there are two possible outcomes to the inclusion of such samples in the calibration and validation sample sets. One is that that samples will be unsuitable for calibrating and validating the NIR procedure. This outcome is detrimental.

The other possible outcome is that the range of variability of the sample set will be increased, resulting in a more robust calibration. This outcome is beneficial. It is usually not obvious a priori, however, which outcome will occur in a given case.

Procedures commonly used for other analytical methods, such as 'spiking' samples, are generally unsuited for use with NIR analysis. While occasional successes are reported, a problem that arises is that the resulting material will again not behave sufficiently closely to 'real' production samples. Augmenting the samples collected from the production process with 'development' samples, made either in the laboratory or in a pilot plant, can extend the range but here, too, such samples are usually not found to be from the same statistical population as the production samples. As with process upset samples, it is usually not clear whether the extended sample set will have a beneficial or detrimental effect on the calibration. None of these proposals, therefore, are completely satisfactory although of necessity, one or another is often used.

Moffat et al. propose to reduce the difficulty of meeting the range requirement and minimize the effect of including non-production calibration samples by reducing the required range to twice the permitted range for the analyte in that product. While following this proposal may reduce some problems, it may introduce or exacerbate others. Reducing the calibration range also affects, usually detrimentally, the quality of the calibration and the values of the auxiliary statistical diagnostics that are calculated to evaluate the calibration performance. The range, correlation coefficient, S.E.E., linearity and other characteristics are related through the properties of the calibration data, so that changing one of these values will change the others. A particular relationship to note is that with all other conditions held constant, reducing the range will reduce the correlation coefficient associated with the regression line that the ICH recommends for testing linearity (see next section), making this, and the other statistics from that regression line, less valuable as indicators of linearity.

Currently the best one can do is to augment the available process samples with development samples and try to obtain development samples that are as similar to the process samples as possible. There is a benefit to this procedure, as well. In addition to increasing the range of analyte, the development samples are likely to exhibit variability in characteristics other than the analyte concentration, making the calibration model more resistant (more robust) to those types of effects of sample variability.

### 2.5. Linearity

In conventional analytical technologies, linearity is often assessed by analyzing samples ranging from 0 to 100% of the product target value. In NIR analysis, this approach sometimes fails, if the calibration model is valid only over a relatively limited range. Over a limited range, assessing the linearity is much more difficult than over the wider range that other technologies enjoy. This situation is exacerbated by the well-known mathematical dictum that any continuous curve will approximate a straight line as closely as desired; the only requirement for this is that the segment of the curve examined is short enough. In the NIR analysis, this situation is exacerbated further by the random noise of the reference laboratory results superimposed on the data.

Even the ICH recommendation [8] to calculate a regression line and report the $Y$-intercept, slope, correlation coefficient and residual sum of squares does not solve this problem completely, because these quantities include contributions from both non-linearity and random variations of the data. As discussed under 'Range', reducing the range

reduces the effect of non-linearity. Therefore, when both non-linearity and random error of the reference laboratory data are present, the reduced range makes any non-linearity present more difficult to observe, both visually and by examining the correlation coefficient and other statistics reported for the regression line.

A statistic exists, however, that can address these limitations of the other statistics. Since this statistic has not previously been widely applied to either NIR technology or pharmaceutical analysis, it is appropriate to discuss this statistic here.

This statistic is the Durbin–Watson statistic (DW). It is calculated from the formula [13]:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n} (X_i - X_{i-1})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad (1)$$

The key to calculating this statistic is that prior to performing the calculation described by Eq. (1), the data must be put into a suitable order. The Durbin–Watson statistic is then sensitive to serial correlations of the ordered data. While this serial correlation is often thought of in connection with time series, that is only one of its applications.

Draper and Smith discuss the application of DW to the analysis of residuals from a calibration [13] (see page 69, also pages 181–192). While we cannot reproduce their entire discussion here, at the heart of it is the fact that there are many kinds of serial correlation, including linear, quadratic and higher-order. As Draper and Smith show (on page 64), the linear correlation between the residuals from the calibration data and the predicted values using the calibration model is zero. Therefore, if the sample data is ordered according to the analyte values predicted from the calibration model, a statistically-significant value of the Durbin–Watson statistic for the residuals is indicative of high-order serial correlation, i.e. non-linearity.

Draper and Smith also provide tables of DW, and also recommend a test procedure that uses upper and lower bounds, to enable the analyst to conclude that non-linearity either is or is not present or that the test data is inconclusive.

### 2.6. Qualification procedure

The term ‘qualification’ has two meanings. One meaning is the one discussed above in Section 2, when we distinguished between ‘qualification’ (instrument responding to the actual analyte) versus ‘specificity’ (instrument not responding to other components of the sample). In this sense, ‘qualification’ refers to the qualification of the analytical method. Ways are needed to verify that the measurement is responding to the presence and amount of the analyte, rather than to the other components of the sample. The fundamental tenets of spectroscopy provide us with tools to perform this function:

1. Spectra of the active ingredient and of the excipients can be measured to determine the locations and strengths of the absorbance bands.
2. Wavelengths used in the calibration model can be compared with the known bands of the active ingredient and those of the excipients to verify that the bands of the active ingredient are being used.
3. Samples containing different amounts of the active ingredient (either naturally different or ‘spiked’) can be compared to verify that the band strengths increase with concentration.
4. Spectra of the ingredients can be multiplied by their concentrations and those added together to reconstruct the spectrum of the sample, the reconstructed spectrum can be compared with the actual sample spectrum. Some of the tools used for creating the calibration model can also assist here:
5. The wavelengths used for the calibration (for MLR models) or the factors used (for PLS or PCR models) can be examined to ensure they are using the actual spectroscopic information from the analyte.
6. For PLS and PCR calibrations, the coefficients can be plotted and regions of large coefficients compared with the spectrum of the analyte.

Not all these techniques need be applied to every calibration, but there should be enough data

gathered to inspire confidence that the calibration model is in fact based on spectroscopically valid information.

The other meaning of 'qualification' is the verification that a given sample is a valid sample for the calibration model in use, in that it conforms to the properties of the samples upon which the calibration was based. Here it is the individual sample that is being 'qualified' for analysis. In this meaning of the term, it is the overall characteristics of individual samples that must lie within some range of variability of the calibration samples. These characteristics are reflected in the overall spectrum of each individual sample, which are then compared with the spectra of the calibration samples.

NIR spectroscopy is notorious for having the property that, especially in complicated mixtures, the absorbance bands of the components of the mixture tend to overlap and interfere with each other; indeed, this is the main reason that the complicated mathematical techniques used, collectively called 'chemometrics', are needed to sort out the spectral interferences and provide accurate analyses. However, NIR instruments enjoy a superb capability to accurately and precisely measure absorbance. Therefore, we use this capability to achieve sample qualification. This can be done because accurate absorbance measurements provide a handle to a means to automate the process of spectral matching. Thus, it becomes possible to accurately compare the spectrum of an unknown sample to the spectral signatures of the samples upon which the calibration was based and for which the calibration model will provide accurate results.

Samples of a given type cluster together in the mathematical multidimensional space in which chemometric calculations are performed [14]. Clusters may be examined visually to see if a sample to be analyzed falls within the cluster defined by the calibration samples for that analysis. Alternatively, statistical measures, such as, Mahalanobis Distance [15,16] may be used to automatically and objectively qualify a sample by determining whether the sample to be analyzed falls within the appropriate cluster.

### 2.7. Specificity

As described above, due to the nature of NIR spectra, the a priori identification of an unknown material based solely on its NIR spectrum has historically not been considered a method of choice. The sample qualification procedure described above is sensitive enough however, to reject samples, for example, that fail to contain any analyte, especially if the correct analyte concentration is high.

For method qualification, on the other hand, we note that the spectral characteristics of pure materials tend to be indicative of the material. This connection between the material and its spectrum can be enhanced by appropriate transforms of the data, for example, first or second spectral derivatives, that is spectra of $dA/d\lambda$ or $d^2A/d\lambda^2$. An appropriate mathematical transform of the spectrum, followed by a computerized spectral matching procedure can verify the identity of the analyte.

On the other hand, similar molecules tend to have similar NIR spectra. Thus, when the analyte is incorporated into a complicated pharmaceutical preparation, the interference and masking of the absorbance bands of the analyte by the other materials in the samples precludes guaranteeing that only the proper analyte is present. Moffat's review discusses this situation at length, and recommend procedures and criteria for verifying whether NIR measurements alone provide sufficient specificity for a given product.

The ICH guidelines provide for the situation where a single analytical technique alone does not provide sufficient specificity, and recommends that when required, the results from two or more techniques can be combined to achieve the necessary level of discrimination. In practice, this can be achieved by use of another, validated method such as mid-infrared spectroscopy. The mid-infrared spectrum of a product will typically be measured before the product is released from the production department for final assay, and this spectrum can be used to verify the identity of the analyte.

### 2.8. Robustness

To some extent, the evaluation of robustness of an NIR analysis overlaps that of intermediate precision, both characteristics concern themselves with the potential effect of extraneous phenomena on the analytical answer. One way to form a distinction between them is the ability to control the extraneous phenomenon; intermediate precision measures the effect of phenomena that cannot be controlled: the possible difference between analyses performed on different days, for example. Furthermore, one cannot go back and redo an analysis performed on a previous day, under the conditions existing that day (conditions that may include the proverbial 'phase of the moon').

Robustness, on the other hand, concerns itself with the effects due to extraneous phenomena that can be controlled, and the guidelines direct that when necessary, such control must be imposed on those phenomena that might affect the answer. Examples of such effects include environmental temperature, humidity, and the like. In a well-controlled laboratory environment, however, these variables may be very tightly controlled and therefore, not vary measurably. In such a case the effect of a variable cannot be ascertained, nor does it need to be ascertained, as long as that variable continues to be maintained constant.

The purpose of determining robustness is to ascertain which variables affect the analytical results, in order that those variables can be controlled when the analytical method is adopted for routine use. Therefore, it is up to the scientist developing the method to determine which of the extraneous variables need to be tested. We again use our own work to provide an example of how this can be introduced into the protocol, for a somewhat non-standard variable. In the current study, reported in the companion paper [11], the samples in the study for which this protocol is being developed have some asymmetry, Both sample types are in the form of elongated objects. Hence, one random source of variability that can exist, and which might affect the readings, is the orientation of the samples. Therefore, all samples were presented to the instrument using a close-fitting holder that allowed no room for the sample to wiggle in the holder. The remaining source of variability was which surface of the sample was presented to the instrument during the measurement. To ascertain this residual effect of orientation, this variable was studied by turning the sample over in the holder, so that a different surface would be exposed to the instrument's optics.

As described above, robustness is not always evaluated at the time the method is developed, but rather is assessed afterward. For example, Moffat et al. did not assess robustness in the report of their study. In our companion paper, in addition to studying the effect of sample orientation, we examined the auxiliary statistics associated with the calibration model. As mentioned earlier, during the course of the calibration calculations, several auxiliary statistics are also computed, which indicate the quality of the calibration. Some of them, such as the S.E.E., are directly related to the performance of the calibration model. Others are less directly related, but have other useful properties. In particular, the $F$ statistic is an indicator of the quality of the other statistics, in so far as it provides a relative measure of the ability of a given calibration model to provide a result consistent with the S.E.E., etc., when measuring samples not included in the calibration data [17]. Thus, the $F$ statistic is a measure of the relative robustness of different calibration models, and we report the $F$-value for the calibration models in addition to the other statistics recommended by the ICH guidelines.

## 3. Conclusions

The unique characteristics of a chemometrics-based analytical method such as NIR spectroscopy were not considered during the development of the FDA and ICH guidelines. Nevertheless, it is possible to devise calibration and validation protocols that enable the guidelines to conform to the specialized requirements of this method. This opens the door for regulatory approval of NIR methods for pharmaceutical analysis.

Some of the required characteristics of evaluation, such as, accuracy and repeatability, can be applied directly just as with any other analytical method.

For some of the characteristics, such as linearity, the novel application of a standard, although uncommon, statistic formalizes the test procedure. Use of a standard, well-characterized statistic puts the test on a more objective basis, and improves the match between the test of the characteristic and the nature of the data.

For one of the characteristics, specificity, where the NIR method is weak, a protocol for testing the suitability of NIR for the particular analyte is available, and if the tests reveal NIR to be not suitable, alternative methods can be used to satisfy that particular requirement.

## References

[1] D. Boyd, B. Maguire, Near Infrared Spectroscopy: Proceedings of the 9th International Conference, Verona, Italy, 1999, NIR Publications, Charlton, Chichester, West Sussex, UK, 2000, pp. 357–363.
[2] H. Mark, G. Kemeny, Chemical Processing 54 (2) (1991) 54–58.
[3] G.C., Marten, J.S., Shenk, F.E., Barton, Near Infrared Reflectance Spectroscopy (NIRS): Analysis of Forage Quality, US Department of Agriculture, Washington, DC, vol. Agriculture Handbook ≠643, 1989, pp.110.
[4] Pharmacopeial Forum, 24 (4), (1998) 6463–6473.
[5] ⟨1119⟩ Pharmacopoeial Forum; 26 (1), Jan–Feb (2000) 237–247.
[6] Physical and Physiochemical Methods, Near-Infrared Spectrometry, European Pharmacopoeia, 1997, pp. 43–44.
[7] ICH-Q2A, Guideline for Industry: Text on Validation of Analytical Procedures, 1995.
[8] ICH-Q2B, Guidance for Industry: Q2B Validation of Analytical Procedures: Methodology, Food and Drug Administration, Rockville, Maryland, 1996.
[9] A.C. Moffat, A.D. Trafford, R.D. Jee, P. Graham, Analyst 125 (2000) 1341–1351.
[10] ASTM, Standard Practices for Infrared, Multivariate, Quatitative Analysis, in: Annual Book of Standards, vol. 3.06, American Society for Testing and Materials, West Conshohocken, PA, 1997.
[11] G.E., Ritchie, R.W., Roller, E.W., Ciurczak, H., Mark, C., Tso, S.A., MacDonald, 2001, in preparation.
[12] ⟨1225⟩ Validation of Compendial Methods, US Pharmacopoeia, vol. 24, 1999, pp. 2149.
[13] N. Draper, H. Smith, Applied Regression Analysis, third ed., Wiley, New York, 1998.
[14] M. Sharaf, D. Illman, B. Kowalski, Chemometrics, Wiley, New York, 1986.
[15] H.L. Mark, D. Tunnell, Analytical Chemistry 57 (7) (1985) 1449–1456.
[16] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: a Textbook, Elsevier, Amsterdam, 1988.
[17] H. Mark, Principles and Practice of Spectroscopic Calibration, Wiley, New York, 1991.