Laboratory test method validation

J.H. LUMSDEN

Department of Pathobiology, University of Guelph, Guelph, ON N1G 2W1, Canada

SUMMARY

Intelligent use and interpretation of any test procedure requires having knowledge of the test reliability in specific clinical situations. For laboratory tests relating to specific diseases, clinical interpretations are based opti- mally upon positive and negative predictive values, or odds ratios, predetermined at useful medical decision limits. Due to the many species, management and disease differences encountered in veterinary medicine, the interpretation of routine laboratory test values is usually made in relation to reference intervals determined for a defined species subset, or to other decision limits dependent upon the experience of the clinician.

Introduction of any new procedure, instrument or reagent is based upon several features including anticipated clinical value and efficiency in a diagnostic laboratory environment. The procedures for validation of a new test in the laboratory are well described. The reasons for each validation procedure and the interpretation as to whether the resulting observations indicate likely ability of the test to meet clinical needs is less well described and understood. If the clinical requirements of a test are the determining criteria validation procedures are much easier to understand and interpret. At present, most veterinary laboratory tests must refer to the recommended clinical requirements for human diagnostic testing.

The experimental plan for within-laboratory method validation is presented in four phases including initial familiarization with the method, preliminary and more extensive validation experiments and implementation. Experimental designs are reviewed briefly for linearity studies, recovery studies, interference studies, within-run, between-run and replication studies, comparison of method studies and reference intervals. Data analyses including the requirements for various statistical tests are described.

If within-laboratory validation experiments indicate likely acceptable clinical performance the test procedure can be implemented for initial clinical use. For tests which relate to a specific disease, prospective studies should be designed, in consultation with clinicians, to evaluate medical decision limits leading to determination of diagnostic sensitivity, specificity and predictive values of positive and negative test values. If clinical interpretation is dependent upon reference intervals, these should be determined according to recommended procedures and the clinicians should be informed as to the source and reliability.

KEY-WORDS: laboratory test - method validation - data analyses.

RÉSUMÉ

Validation des analyses de biologie médicale. Par J.H. LUMSDEN.

L’utilisation rationnelle et l’interprétation de tout test de laboratoire nécessite une connaissance préalable de la fiabilité du test dans des situations cliniques. Pour des tests de laboratoire de diagnostic de maladies spécifiques, les interprétations cliniques sont fondées de manière optimale sur des valeurs prédictives positive et négative, des rapports de cotes (odds ratios), prédéterminés pour des limites de décision utiles en médecine. En raison de la variété des espèces, des conditions d’élevage et des maladies rencontrées en médecine vétérinaire, l’interprétation des tests de laboratoire de routine est généralement faite en relation avec les intervalles de référence déterminés dans une espèce donnée, ou bien en fonction d’autres limites de décision dépendant de l’expérience du clinicien.

L’introduction d’une nouvelle procédure, d’un nouveau matériel ou de nouveaux réactifs repose sur différents critères, dont la valeur clinique espérée et l’efficacité dans l’environnement du laboratoire. Les procédures de validation de nouveaux tests sont bien décrites. Les motifs de chaque procédure de validation et d’interprétation, c’est-à-dire la probabilité que les résultats du test répondent aux besoins de la clinique, sont moins bien compris et moins bien décrits. Si les besoins cliniques sont les critères essentiels, les procédures de validation sont plus faciles à comprendre et interpréter. Actuellement, la plupart des tests de laboratoire se réfèrent aux besoins cliniques recommandés pour le diagnostic humain.

Le plan expérimental de validation de méthode intra-laboratoire en quatre phases comprend une phase initiale de familiarisation, puis des étapes de validation préliminaires et plus approfondies, avant la mise en service. Les modalités expérimentales sont passées en revue rapidement pour l’étude de la linéarité, de la récupération, des interférences, de la précision intra- et inter-séries, de la comparaison de méthodes et des intervalles de référence. L’analyse des données et les besoins pour effectuer les tests statistiques sont indiqués.

Si les expériences de validation intra-laboratoire indiquent des performances vraisemblablement acceptables au plan clinique, le test peut être mis en service pour une utilisation clinique initiale. Pour les tests visant une affection particulière, des études prospectives devraient être mises au point, en liaison avec les cliniciens, pour évaluer les limites de décision médicale conduisant aux déterminations de la sensibilité, de la spécificité diagnostiques et des valeurs prédictives positives et négatives. Si l’interprétation clinique dépend de l’intervalle de référence, celui-ci devrait être déterminé avec des procédures recommandées et les cliniciens devraient être informés de la source et de sa fiabilité.

MOTS-CLÉS: analyse de biologie médicale - validation de méthode - analyse de données.
Introduction

All laboratory tests must be validated before being introduced for patient testing to insure that the values reported will meet clinical expectations with a desired degree of reliability. Re-validation is required, to a less or greater degree, following any change in reagents, instrumentation or protocol. Method validation begins with the considerations for, and selection of, a new test method for introduction into the laboratory or for patient-side use. Evaluation and validation of the method analytical performance is required to assess the degree of error expected due to inaccuracy and imprecision and to confirm that the degree of error meets the anticipated clinical requirements. The clinical requirements should be predetermined prior to initiating method validation. The procedures recommended for method validation differ with the type of test and the anticipated use. Recommendations frequently differ between authors and sources of information. Experiments must be designed so that the correct data are obtained. The appropriate statistical tools must be used to correctly estimate errors. Statistical tests, although necessary, do not prove acceptability. The final objective decision for acceptance of a method must consider error assessment, practical technical and financial aspects as well as the anticipated ability to meet clinical requirements.

Validation procedures are more demanding for a method developed within the laboratory than for one developed by a manufacturer. Most manufacturers provide some performance data which can be used for comparison when assessing initial performance within the laboratory of interest. Tests using complex methodology usually require more validation steps than moderately complex or simple methods. Point-of-care methods require some validation in each laboratory or clinic setting, especially when used for species other than those for which the methods were originally designed.

The objectives, procedures and the study designs for method validation are reviewed. The graphical and statistical evaluations of method comparison studies are compared. Validation of diagnostic performance is the final step in method validation and requires clinical cooperation and application of epidemiological principles.

Selection of a new or revised method

New test procedures are considered for a variety of reasons. Research reports often initiate interest by clinicians or instrument and reagent manufacturers. New instruments and reagents may appear to provide improved accuracy or precision, to reduce technical effort, to improve efficiency, to reduce reporting time and/or cost per test.

Method selection should begin with a clinical perspective. The primary consideration should be the ability to produce results which have sufficient analytical reproducibility and accuracy to meet anticipated clinical requirements. This decision should be based upon examination of all available data estimating error of the method and insuring that the observed error is acceptable for diagnostic purposes.

Candidate methods must meet the practical requirements of the laboratory. Are there adequate space, equipment, and personnel? Laboratory personnel with the aid of technical literature can assess most practical requirements. Discussions with clinicians may be required to estimate throughput and whether turnaround will meet desired reporting time. The estimated cost per test must consider the level of anticipated use. Is there adequate supporting literature to expect the test to be used for a reasonable time or is this a current 'hot' research topic with minimal clinical validation or experience? Economic aspects should be weighed only after practical considerations appear attainable.

Evaluating and validating methods

The reasons for using various procedures when validating laboratory methods is not always made clear in scientific publications. Similarly, the correct use and the interpretation of statistical tests and the objective interpretation of acceptability for laboratory tests being validated is not always apparent. I was encouraged to read the description by Dr. WESTGARD as to his early experiences (http://www.westgard.com/essay15htm), since it mirrors my observations. Most of the uncertainty regarding the decision as to method acceptability is alleviated if one understands that the 'inner, hidden, deeper, secret meaning of method validation' is 'error assessment' (www.westgard.com/essay15htm). How much error might be present in the test result within your laboratory? Could this degree of error affect the interpretation and possibly patient care? If the potential error is large enough to lead to misinterpretation, then the method is not acceptable.

The method validation process becomes much easier to understand if the focus is directed towards the sources of potential analytical errors and how these errors can be investigated. What experiments, and which experimental designs, will best demonstrate the errors? How many observations are required to obtain good estimates of the error? For the experimental design which statistical method should be used to describe the extent of the error observed? How much error can be accepted within a method without affecting interpretation? This latter question requires consultation with clinicians, if possible, and should be predetermined.

ERROR

Analytical (total) error is the summation of random error (imprecision) and systemic error (inaccuracy or bias) [17]. Random error is the amount of variation inherent in the method. Systemic error is the difference from the true value. Although total error is the important question for the clinician, random and systemic errors are usually examined independently in the laboratory.

Random error is both positive and negative relative to the observed mean value of replicate determinations [17]. These replicate observations from a single sample can be plotted as a histogram to illustrate the spread around the mean (m). The calculated standard deviation (SD) is used to quantitate the degree of imprecision. This imprecision can be expressed in percentage as the coefficient of variation (CV) using the formula CV = SD x 100 / m. For methods where the observed
SD increases as the analyte concentration or activity increases, the CV remains relatively constant. If the SD remains constant as the analyte concentration or activity increases the calculated CV will be lower at higher analyte concentrations. The opposite might also occur. Thus, the analyte concentration or activity must be known for the CV to have relevance. The CV provides ready comparison of method imprecision at different analyte concentrations or for comparison between analyte methods since it is expressed as a percentage and is independent of the concentration or activity units involved.

Systemic error (bias) may be constant or proportional and is either positive or negative, as compared to the true (correct) value [17]. Comparison of methods is used to assess systemic error. The bias is calculated as the average difference, or the difference between averages, for observations obtained using both the ‘new’ and the ‘old’ or a ‘reference’ method.

**ACCURACY**

Accuracy of a method is defined by the International Federation of Clinical Chemists (IFCC) as the closeness of the agreement between the measured value and the "true" value [17]. The accuracy of a new method can be described in terms of either the systemic error or the total error relative to the best available estimate of the "true" value. Definitive methods, such as mass spectrometry, are used to develop primary reference materials which can then be used for development of reference methods by manufacturers [17]. Comparative method means have been shown to closely approximate true values. Comparative method means, to a great extent, are obtained from the observations generated by multiple laboratories using a variety of instruments and techniques. Comparative method means have replaced use of reference laboratory means [17]. Peer group means are obtained from the proficiency testing results of several laboratories using similar instruments and techniques. The manufacturer is relied upon to have made comparison to some independent measure of accuracy.

Many national and international committees continue to develop and provide information relating to the accuracy base for analytes of clinical interest. Increasingly this information can be obtained at various sites on the World Wide Web (e.g. [http://www.aacc.org](http://www.aacc.org) for the Standards Committee of the American Association of Clinical Chemistry). For practical reasons, various sources of sera are used as estimates of the "true" value in veterinary diagnostic laboratories.

**Clinical requirements**

In order to be able to assess whether a new method is able to meet expectations, the clinical requirements must be defined. Criteria for test method clinical requirements are usually related to the biological distribution of values observed within a healthy population [3, 7, 8, 9, 36, 39] or to consensus opinions based upon perceived requirements for clinical diagnosis [35, 40]. Tonk’s ‘Allowable Limits of Error’ (ALE) originally proposed in 1960 were based upon the relationship of imprecision to the reference interval and expressed in percentage [39] using the formula ALE (%) = ±1/4 normal range x 100 / mean of normal range. The original maximum ALE of 10 % was increased to 20 % due to the inability of many laboratory methods to reach the original objective.

The current allowable error recommendations are listed for routine tests including clinical chemistry, hematology, endocrinology and related markers, for immunology and for toxicology and therapeutic drug monitoring [17]. The Clinical Laboratory Improvements Amendments (CLIA’88) guidelines [40] for acceptable analytical performance can be viewed at ([www.westgard.com/clia.htm](http://www.westgard.com/clia.htm)). The European Calculated Biological Allowable Total Errors can be viewed at ([www.westgard.com/europe.htm](http://www.westgard.com/europe.htm)).

Examples of acceptable analytical performance [17, 40] appear in Table I. Until species specific data is available veterinary diagnostic laboratories must rely upon criteria defined for medical laboratories.

<table>
<thead>
<tr>
<th>Test or Analyte</th>
<th>Acceptable Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine aminotransferase</td>
<td>Target value ± 20%</td>
</tr>
<tr>
<td>Albumin</td>
<td>Target value ± 10%</td>
</tr>
<tr>
<td>Alkaline phosphatase</td>
<td>Target value ± 10%</td>
</tr>
<tr>
<td>Amylase</td>
<td>Target value ± 10%</td>
</tr>
<tr>
<td>Bilirubin, total</td>
<td>Target value ± 6.8 µmol/l or ± 20% (greater)</td>
</tr>
<tr>
<td>Blood gas pO2</td>
<td>Target value ± 3 SD</td>
</tr>
<tr>
<td>Blood gas pCO2</td>
<td>Target value ± 5 mm Hg or ± 8% (greater)</td>
</tr>
<tr>
<td>Calcium</td>
<td>Target value ± 0.2495 mmol/l</td>
</tr>
<tr>
<td>Chloride</td>
<td>Target value ± 5%</td>
</tr>
<tr>
<td>Creatinine</td>
<td>Target value ± 26 µmol/l or ± 15% (greater)</td>
</tr>
<tr>
<td>Glucose</td>
<td>Target value ± 0.33 mmol/l or ± 10% (greater)</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Target value ± 25%</td>
</tr>
<tr>
<td>Potassium</td>
<td>Target value ± 0.5 mmol/l</td>
</tr>
<tr>
<td>Sodium</td>
<td>Target value ± 4 mmol/l</td>
</tr>
<tr>
<td>Total protein</td>
<td>Target value ± 10%</td>
</tr>
<tr>
<td>Leukocyte count</td>
<td>Target value ± 15%</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>Target value ± 6%</td>
</tr>
<tr>
<td>Platelet count</td>
<td>Target value ± 25%</td>
</tr>
<tr>
<td>Cortisol</td>
<td>Target value ± 25%</td>
</tr>
<tr>
<td>Thyroid stimulating hormone</td>
<td>Target value ± 3 SD</td>
</tr>
<tr>
<td>Thyroxine</td>
<td>Target value ± 13 mmol/l or ± 20% (greater)</td>
</tr>
</tbody>
</table>

**Experimental plan for validation of a new method**

Four phases have been described including the initial familiarization with the method, preliminary validation experiments, more extensive evaluation of precision and accuracy and implementation for routine use [17]. The preliminary validation includes those steps that can be done more easily within a few days. If unsatisfactory, a decision may be made...
that the method is unacceptable. More extensive precision and validation evaluations must be studied over a minimum of 20 working days.

**FAMILIARIZATION**

The familiarization phase includes establishment of the working procedure, checking the working range and the calibration. The detection limit may be determined initially or in later experiments. Establishment of the working procedure includes preparing reagents, setting up the instrument, calibrating the methods and obtaining results from test samples. The standards must be carefully checked to insure that the calibration is correct.

**Linearity studies**

The desired working range is defined during the preliminary evaluation, or is assessed from the manufacturer’s specifications for the method. The working range can be validated as part of the familiarization studies. Two pools of sera are selected, one with analyte concentration or activity close to zero, or the detection limit, and the other with a high concentration, or close to the expected upper limit of the working range. Varying proportions of sera are mixed to create an additional three, or more, concentrations, thus at minimum five specimen levels. Two [17], three [43] or four [24] measurements are made for each specimen. The observed values are plotted on the x axis against absorbance on the y axis. Visual inspection for linearity is used as an estimation of the working range. Varying proportions of sera are mixed to create an additional three, or more, concentrations, thus at minimum five specimen levels. Two [17], three [43] or four [24] measurements are made for each specimen. The observed values are plotted on the x axis against absorbance on the y axis. Visual inspection for linearity is used as an estimation of upper working range for the method. At the same time the absorbance of the zero blank can be assessed as to potential significance for the method.

**PRELIMINARY VALIDATION**

The preliminary validation experiments include within-run replication, interference, recovery and judgement of analytical acceptability [17].

**Within-run replication**

The initial replication study assesses random error within-run or within-day. Samples are chosen which approximate medical decision levels of greater interest for the test. The matrix of the sample, whether serum, urine, cavity fluid, etc., should approximate the matrix of clinical specimens for the analyte of interest. Control solutions made with a similar matrix as patient samples, e.g. lyophilized serum, have pre-determined analyte concentrations as well as long term stability. Aqueous solutions may be used for preliminary studies of some analytes. Pools of fresh patient samples are used frequently for initial within-run / within-day replication studies. Two or three samples are selected with analyte concentrations which approximate important medical decision limits [16, 17].

Minimums of 20 replicates are recommended for each sample. The mean, SD and CV are calculated and compared to allowable SD, or CV, for the method. Within-run SD’s should be about 1/2 the allowable SD since additional sources of error are usually observed for between-run precise studies [43]. The SD of a method can be calculated from the values observed for patient samples done in duplicate where the SD equals the square root of (the sum of the differences squared divided by two times the number of samples).

**Interference studies**

Interference studies are done initially to assess the effects of hemolysis, hyperbilirubinemia and hyperlipemia [10, 14, 15, 25]. Other interferences may be studied at a later time [45]. The interference studies provide information about potential systemic error as may be caused by lack of specificity of the method. Different experimental designs are described. If a method is available which is known to be free of such interferences a series of samples containing increased concentrations of the interferent are analyzed using both methods and the results compared [17]. Test samples may be prepared by adding interfering materials to one aliquot and an accurately pipetted equal volume of solvent or water to the other aliquot. The analyte concentration is tested in each aliquot, preferably in duplicate or triplicate, and the observed values are compared for interference effects [10]. Increasing amounts of the interferent can be added to pooled species sera. Dilution effects must be considered when assessing results. The average difference between aliquots with, and without, added interferent can be calculated and plotted. The difference, or error, when the interference is present is compared to the error that is allowable for the test. The method is assessed as acceptable or unacceptable for the analyte for the species.

Due to the species differences reported, for each method interference experiments must be done for each species of interest for veterinary diagnostic laboratories [14]. Visual assessment can be observed from graphical presentation. When automatic correction factors were developed for hemolysis, bilirubinemia and lipemia in an automated instrument designed for use with human sera but used with domestic animal species sera, the interspecies differences precluded reporting patient values when these interferences were noted (personal observations). Significant effort is required to develop allowable limits for the three common interferences in each animal species sera. In veterinary diagnostic laboratories, at minimum, the effects of these common interferences should be studied to determine whether there is a positive or negative bias. Ideally, the maximum concentration of the interferent that creates error greater than allowable for the method should also be determined.

**Recovery studies**

Recovery studies assess proportional systematic error due to competitive reactions from substances within the specimen, including matrix effects [17]. Different amounts of the substance of interest are added to baseline sera pools, using concentrated solutions, so that dilution effects are less than 10 % [17]. Experimental samples are analyzed in quadruplicate in order to detect small additions or duplicates if the additions are large [17, 44]. The original concentration is subtracted from the final observed concentration to determine the amount recovered. This amount is divided by the amount added, multiplied by 100 and expressed as the percent recovery. The error observed should be less than the allowable total error pre-assigned for the method.
Recovery studies are subject to problems associated with design and performance, calculation of the data and interpretation of the results. If a very reliable comparison method is available, e.g. reference method, a method comparison study should be given higher priority than a recovery study [44]. When bias is observed for the method comparison study, the recovery study may help to explain the bias observed [17, 44].

**DETAILED VALIDATION STUDIES**

Providing that the preliminary validation studies are favorable, more extensive experiments are initiated. These include replication studies conducted over at least 20 working days, comparison of methods studies, judgement of analytical acceptability, verification of reference interval and documentation studies.

**Replication studies**

Replication studies are done to estimate method imprecision due to random error. In addition to the initial studies confirming acceptable within-run precision, between-run experiments are designed to provide a more realistic estimation of overall errors which may occur due to instrument instability, reagent preparation, changes in ambient temperature, analyst, etc. Such estimations are termed "between-day", "day-to-day" or "total" imprecision [26]. Between-day precision should be examined over 20 days. Thus sample stability is a requirement. Lyophilized control sera is often used because if greater stability and because high and low analyte concentrations allow evaluation of performance using concentrations which approximate important medical decision limits. The observed mean, SD and CV are calculated and compared to allowable total error. For within-run or within-day the acceptable SD is 1/4 or less the total error while for between-day studies the SD should be 1/3 or less than the defined total error [17, 40]. Experimental studies of precision can be designed to make use of analyses of variance [26].

**Comparison of methods**

A comparison of methods study is done to estimate systemic error (inaccuracy or bias) within the new method. If bias is present, and if the appropriate statistical calculations are done, systemic error will be identified as constant, proportional, or a combination [17, 44]. The experimental design and the statistical tests used are critical for identification of systemic errors. The primary effort within the laboratory should be devoted to collecting the appropriate data for method validation [43].

The 'new' method is compared to an existing 'comparative' method. Ideally this will be a high quality 'reference' method with documented accuracy so that observed differences can be assigned to the new method. In diagnostic laboratories the new method is compared most frequently to a routine, or 'field' method using patient samples. When observed differences between the methods are small each method is considered to have similar relative accuracy. If observed differences are large, and unacceptable, further studies are required to identify which method is inaccurate.

**Experimental design**

Experimental design must consider the number of samples to be compared, whether single or duplicate analyses will be performed and the time over which the comparison will be made. It is important that the samples are analyzed within a maximum of a few hours so that differences may not be due to sample handling effects [17, 44]. Experimental design and the correct approach to data analyses continue to create discussion [11, 16, 20, 27, 31, 37, 41, 43]. The following recommendations apply to continuous data. If data is not continuous, kappa statistics are designed to allow comparison of categorical data [4, 5].

Most authors, or working groups, recommend comparing a minimum of 40 patient specimens. Selection of specimens with concentrations representing the working range of the method is critical to the success of the comparison. A larger number of specimens, e.g. 100, will increase the probability of detecting interferences due to sample matrix influences and thus whether the specificity of the new method is similar. Ideally, each specimen is tested in duplicate by each method. The samples should be placed in a different order within a run or ideally within different runs if completed within 2 hours. Duplicate testing increases the probability of identifying numerous potential sources of error including valid outliers [17, 44]. Without duplicate testing it is even more important that large method differences are identified immediately so that these specimens can be reanalyzed while still available.

Method comparison should be done over a minimum of 5 days, e.g. 8 specimens a day, by each method within 2 hours [44]. If discrepancies are noted the comparison should be continued for an additional 5 days. Results should be graphed immediately and visually inspected for discrepancies [17, 44]. Where large differences are noted the patient samples should be reanalyzed immediately to eliminate errors due to recording or sample identification.

**Data analysis**

Analysis of data from comparison of method studies continues to remain an actively discussed subject [2, 11, 12, 16, 17, 18, 27, 31, 33, 43] The agreed first step is to plot the observed values and visually examine the data using difference plots and / or comparison line plots. Comparison (line or x/y) plots are made with the comparative method values on the x-axis and the new method values on the y-axis. Visual inspection and drawing a line of best fit provide indication of the analytical range of the data, the linearity and the relationship of the methods as shown by the closeness of the values to the line of identity and the intercept with the y-axis [16].

**Bias plots**

Difference, or bias plots can be made using various values for the x-axis and the y-axis [16]. The values to be used are dictated by the degree of reliability of the comparative method and whether duplicate analyses have been determined. If the comparative method is of high quality, such as a reference method or well studied routine method, the comparative method results are plotted on the x-axis and the difference between the new and the comparative method on the y-axis [27]. If duplicate measurements are made, the difference between all measurements and the 'true' value can be plotted allowing estimation of the precision of the replicate measure-
ments [16]. If the comparative method is a routine method of unconfirmed reliability, the average of the comparative and new method observations for a patient sample is plotted on the x-axis [2].

When visually assessing difference plots the observations should distribute evenly above and below the zero line over the range of observations [16]. If not evenly distributed above or below the zero line there is likely a bias. If the distribution changes with concentration, constant or proportional errors may be apparent. As the concentration increases changes in precision may be observed, especially if the observations include duplicate values [17].

**Statistical analysis**

Statistical analyses are used to provide numbers which can be incorporated into the overall considerations as to whether a method is acceptable. The use of statistical tests is easier to understand if detection of errors within the ‘new’ method is the objective for using statistical tests to examine the method comparison observations [43]. The statistical tests are not used to prove test acceptability [16, 43]. Also, it is useful to know that the recommended statistical tests for assessing agreement may not be the correct tests to use for prediction of medical decision limits including reference intervals [16].

The traditional statistical approach for analyzing comparison of methods studies is to use linear regression to estimate the slope and intercept and a ‘t-test’ of the mean observations to detect bias. When using simple linear regression it is assumed that the comparative method is free from error, that error in the ‘new’ method is normally distributed and that the error is constant over the sampling range [16]. If the patient samples chosen for study are inadequate, e.g. due to including too narrow a range of observations, and unless these initial statistical tests are used correctly there may be significant misinterpretation as to acceptability of the method [37]. The 40 to 100 samples traditionally recommended for method comparison studies may be inadequate if the range of observations is narrow [20], as for electrolytes where the range observed in healthy individuals may be wide (potassium) or narrow (sodium) relative to the average concentration, i.e. range ratio. For a range ratio of 2, the author states that 544 samples are required to detect one standardized slope deviation whereas if the ratio is 10, then 64 samples can attain the desired differences of medical importance [20].

Simple linear regression provides a good estimate of slope and intercept if the correlation coefficient (r) is 0.99 or larger [27, 37]. If r is less than 0.975, the range of data may be inadequate and simple linear regression method may not be able to provide a good estimate of slope and intercept [27, 37]. If this applies, improvement of the data range or use of alternate statistical methods is required [27]. Visual examination of plotted observations should allow estimation as to whether the imprecision is similar for both methods and whether there is a change with sample concentration. Examination of residual plots from regression analysis can be useful for making this assessment [27]. It is important to note that although the slope of the line for simple linear regression may be ideal, i.e. r = 1 using Pearson’s product correlation, there may be significant constant bias or marked imprecision [33]. Similarly, when using t-test statistics there may be no difference in mean observations even when there is a large proportional error. The t-test statistic is mainly useful for indicating if sufficient data has been collected to reliably estimate bias [43].

For the values observed from many comparative studies alternative statistical tests should be used. These more complex statistical tests are now readily available commercially and from the World Wide Web [www@analyse-it.com; www.westgard.com]. Many of these programs provide ‘help’ sites that include background information detailing use and interpretation of the tests. The current debate within the scientific literature and the evolution and refinement of statistical programs should lead to increased agreement in opinions encountered by readers. One of the following statistical tests should be used if the assumptions required for using simple linear regression have not been met.

The DEMING regression method allows for imprecision in both methods when assessing the degree of agreement 16]. Single or duplicate observations can be analyzed and plotted to provide statistical and visual demonstration of error including imprecision, bias and whether constant or proportional [www@analyse-it.com]. According to a recent report [23], iteratively reweighted general Deming regression can be used to produce statistically unbiased estimates of systemic bias and reliable confidence intervals of bias for all cases.

The PASSING & BABLOK regression method allows for imprecision in both methods, the imprecision does not have to be normally distributed nor constant over the range examined and extreme values can be included [16, www@analyse-it.com].

Concordance correlation coefficient [18] has been proposed as an alternative method to avoid the deficiencies of the paired t-test and PEARSON’S product correlation [33]. The concordance correlation coefficient is used to calculate a number which categorizes test performance as good to poor. This approach should be considered where allowable total error has not been predetermined.

**Acceptable performance**

Judgement as to the acceptability of a ‘new’ method requires consideration of the study design, the reliability of the comparative method and the visual and statistical examination of method comparison data. The method total error is calculated as the measured error (bias) + 3 x the measured SD, i.e. systemic plus random error [41], or systemic error + 4 x SD [17], or total error = 1.65 (imprecision) + inaccuracy [13]. The calculated total error must not exceed the predetermined allowable total error.

A method evaluation decision chart has been designed to incorporate measured bias, measured imprecision and allowable total error [42, www@westgard.com]. The location of the ‘operating point’ is used to judge method acceptability [42, 44].

The use of scatter and difference plots and the graphical and statistical interpretation of acceptability of a new (field) method versus a reference method have been addressed [12]. Difference plots allow examination and interpretation of the
new method observations according to specific criteria. The standard deviation of the differences in comparative method studies is considered to be an indispensable tool for evaluation of aberrant-sample bias (matrix effects) [12].

**Reference intervals**

Reference intervals, and other medical decision limits, must be validated for each method used in a laboratory [40]. If the reference intervals were reliable for the comparative method and there is acceptable agreement between methods, the method comparison data is used to predict reference intervals for the new method [17].

If there is doubt about the reliability of the previous limits, or if the comparison data do not allow confident estimation of limits for the new method, such as due to apparent matrix effects, *de novo* reference limits should be determined [17, 21, 22, 34].

The method comparison data and the types of errors observed dictate the recommended statistical test for estimation of medical decision limits for a new method. Prediction of limits is done using statistical programs which are similar to those used for agreement, but which incorporate different algorithms [16, www@analyse-it.com].

Linear regression, using least-squares estimation, requires that the relationship is linear, that the comparative method is free from imprecision and that measurement error for the new method is normally distributed and constant over the sampling range [16]. If these assumptions are not met alternative methods should be used for predicting reference or medical decision limits [1].

**DEMING regression** has been widely adopted in clinical medicine because imprecision can present in both methods [6]. The imprecision should be normally distributed [38]. If the imprecision is not constant over the sampling range the weighted DEMING regression [19] is recommended.

**PASSING & BABLOK regression** is recommended when imprecision occurs in both the comparative and the new method data [30]. The imprecision need not be normally distributed nor have constant variance over the sampling range. The primary restriction is that the ratio of the imprecision x/y must be equal to the slope squared [29]. Extreme values do not unduly affect the regression line [30].

**IMPLEMENTATION PHASE**

During the implementation phase the method protocol is written in a format that can be used to train new analysts and to meet standard operating procedure requirements [17]. The quality control procedure is chosen and implemented. The method is introduced for service. The clinicians should be informed as to the availability of the method, anticipated scheduling, precision at important medical decision limits, the reliability of reference limits, i.e., interim or final. It is important to monitor performance closely for the first month with regular reviews for several months. Sources of problems are identified, preventive maintenance procedures are improved and quality control procedures updated [17].

**Point of care instrumentation**

It is anticipated that 'point of care' testing will increase significantly, in wards, in intensive care units and for field use [28]. There are too few reports documenting validation of these 'point of care' methods when used for various animal species samples. Similarly, method comparisons with central laboratory methodology is seldom determined or provided to users. Critical care clinicians, technologists and nurses require further knowledge about method validation. Manufacturers specifications are often limited to study of human specimens for accuracy, precision and reference limits. The comparison of methods used for point of care tests presents challenges similar to comparison of method studies for tests used in the central laboratory and in critical care or for 'out-of-hours' testing. Do the tests provide similar or different values? Are there differences in medical decision limits, reference limits, imprecision? Should separate reference intervals be used? How should this data be electronically filed for future use?

What are the minimum steps required for validation of point of care methods for use with animal species? Without published and accepted guidelines one can only suggest adherence to the basic principles discussed previously. The intent is to detect those 'errors' which might significantly affect clinical decisions. Method selection requires careful review of manufacturer's information and method specifications. How can studies be designed that will detect likely sources of error and remain practical and economical? Has the manufacturer, or other users, done method comparison studies? Is this data available? How has the data been presented and analyzed? Is this information likely transferable? Have reference values been developed using adequate numbers of clinically healthy individuals for the species of interest? Can the purchaser expect assistance from the manufacturer? Each of these questions should be addressed.

**Clinical performance**

The clinical performance is the ultimate assessment of the value of a laboratory test. Clinical performance, as compared to analytical performance, further assesses and describes the ability of a test to assist making clinical decisions (usually diagnoses) under defined circumstances. It has been stated that medical decision limits should be used by clinicians instead of reference intervals [32]. When laboratory tests are being used to rule-in and rule-out disease, or to guide prognosis, decision limits are of great benefit, if determined. When laboratory tests are being used to screen for disease and to assess pathophysiology reference limits are used extensively. Alternatively, knowledge of analytical precision is required when monitoring patients. Only if the new and the previous methods have been compared and validated to have similar analytical performance can medical decision limits, or 'cut points' be transferred to a new test method.

It is not the intent of this manuscript to address validation of clinical performance except to emphasize the importance as a necessary step in method validation. At defined decision
limits diagnostic sensitivity and diagnostic specificity indicate the ability of a test to detect disease in sick individuals and to indicate no disease in healthy individuals, respectively [32]. The more likely clinical scenario is to use a test to indicate possibility of disease as well as to differentiate two diseases. Thus, depending upon the clinical question, i.e. how the test is used, the same test will have very different diagnostic sensitivity and specificity. If disease prevalence is known, the predictive value of a positive and a negative test, or likelihood ratios, can be calculated [32].

Some validation of diagnostic performance can be calculated retrospectively, providing a method has been used long enough to accumulate the necessary data. Prospective study designs have many advantages. The primary requisite is a sample size large enough to accumulate the necessary data. Prospective studies, or likelihood ratios, can be calculated [32].

References