ASVCP Quality Assurance and Laboratory Standards Committee (QALS) Guidelines for the Determination of Reference Intervals in Veterinary Species and other related topics: SCOPE

Gräsbeck and Saris introduced the concept of population-based reference values in 1969 to describe the variation in blood analyte concentrations in well characterized groups of healthy individuals (Gräsbeck, 1969). Lumsden and colleagues subsequently began applying these concepts to veterinary species. (Lumsden 1978, Lumsden, Rowe, Mullen 1980, Lumsden, Mullen, Rowe 1980, Lumsden, Mullen, McSherry 1980 and Friendship 1980). Reference values are typically reported as reference intervals (RI) comprising 95% of the healthy population. Since their introduction, population-based RI have become one of the most commonly used laboratory tools employed in the clinical decision-making process (Horn and Pesce, 2005, Chapter 1, page 1-2). Although use of population-based RI is universally accepted, the optimal method for their derivation is frequently debated and is a recurring topic in the clinical laboratory literature. 'Reference interval' is the preferred terminology, and the use of 'reference range' is discouraged.

The standard for production of human population-based RI was commissioned by the International Federation of Clinical Chemistry (IFCC) in 1970. The Expert Panel on the Theory of Reference Values (EPTRV) authored a 6-part series on the production of reference values, which was adopted by several professional organizations including the Clinical and Laboratory Standards Institute (CLSI).(Solberg 1987, PetitClerc 1987, Solberg 1988, Solberg 1991, Solberg 1987, Dybkær 1987) Solberg and Gräsbeck followed with a summary of these recommendations. (1989) Subsequent to these publications, alternative statistical methods for identifying outliers, analyzing reference values, and determining the need for partitioning have been proposed. In addition, concerns were expressed regarding the complexity and expense of compliance with the original IFCC-CLSI standard and the lack of recommendations for handling small reference sample sizes. This led to a revision, completed in 2008, that includes recommendations for transference and validation of RI from other sources and promotes robust methods for determining RI from small sample sizes. (Horowitz, 2008)

The ASVCP has recommended adherence to the CLSI-IFCC guidelines for the determination of population-based RI. However, guidelines specifically addressing veterinary species would have numerous benefits to the veterinary medical community. In response, the QALS Committee of the ASVCP formed a subcommittee to generate guidelines for the determination of reference intervals in veterinary species and to address additional topics of interest. The goal of this subcommittee was to develop balanced and practical recommendations that are statistically and clinically valid. Guidelines that are excessively complex or inaccessible due to high cost or large reference sample sizes will fail to create the desired continuity within the veterinary community. In addition to providing guidelines for the determination of de novo population-based RI for new tests or methods or for new populations of animals, this document provides recommendations on related topics, including transference and validation of RI from other sources, subject-based and common RI, and establishing decision thresholds (or decision limits). Because RI are specific to a particular set of conditions, the document also discusses the misuse of

published RI when reference populations, analyzer and methodology and other pertinent factors are not described. A consistent approach to the development and reporting of population-based RI, and other clinical decision-making values, will benefit all veterinary professionals.

These consensus guidelines are modeled on the revised guidelines of the CLSI-IFCC for establishing population-based RI, which were summarized in Veterinary Clinical Pathology. (Horowitz 2008; Geffre 2009) Recommendations on other topics are based on current literature and on the experience of individuals working in veterinary laboratory medicine. These guidelines were independently reviewed by experts in the field of clinical laboratory medicine and by the ASVCP membership. They were subsequently approved by the QALS committee and the Executive Board of the ASVCP.
As a guideline, these procedures may be applied as written or modified by the user for specific purposes. The intended users of these guidelines include individuals working in veterinary reference diagnostic laboratories, animal research clinical laboratories, manufacturers of veterinary diagnostic equipment and assays, and authors of RI articles in veterinary species. In addition, veterinary clinicians who use RI should be familiar with these procedures so that they can evaluate RI studies to ensure appropriate application to their patient population. A list of definitions for terms used in RI studies can found at the end of this document.

QALS Reference Interval Team
Kristen Friedrichs, chair
Kirstin Barnhart
Julio Blanco
Kathy Freeman
Kendal Harr
Balazs Szladovits
Raquel Walton

Table of contents

Preliminary investigation	p. 4
Selection of the reference population	p. 4
Preanalytical procedures – patient preparation, sample collection and analytical quality assessment	p. 6
Analytical procedures	p. 7
Statistical analysis of reference values	p. 7
Postanalytical procedures – laboratory presentation of reference intervals	p. 13
Transference and validation of reference intervals	p. 14
Common (or multicenter) reference intervals	p. 17
Use of published reference intervals	p. 18
Biological variation, individuality and subject-based reference intervals	p. 19
Establishing decision thresholds (or decision limits)	p. 20
References	p. 23
Figure 1. Procedural steps for *de novo* determination of RI for new methods or new populations.	p. 27
List of definitions	p. 28
Table 1. Criteria for the selection, partitioning or exclusion of reference individuals.	p. 31
Table 2. Recommended procedures for establishing RI based on reference sample size and distribution.	p. 32
Table 3. Information to include in RI study document of when publishing RI studies	p. 33
Tables for reporting reference data when sample size is < 40	Addendum

DETERMINATION OF *DE NOVO* REFERENCE INTERVALS FOR NEW ANALYTES, NEW METHODS OR NEW POPULATIONS

Preliminary investigation

Investigation of sources of biological variability and interference affecting measurement of the analyte(s) in question is recommended in order to determine specifications for collection and handling of samples and for selection and preparation of reference individuals. This information also may be used to establish inclusion and exclusion criteria and determine the need for separate RI based on animal factors (e.g., age, sex, breed) or preanalytical techniques (e.g., serum vs. plasma). Analytical interference from bilirubin, lipemia, and hemolysis may be considered in this investigation; however, reference samples with these alterations typically are rejected as evidence of illness, non-fasting, or poor sample handling.

Selection of the reference population

1. Define the reference population of interest, as well as the criteria used to confirm health in individuals selected from this population (selection, inclusion, exclusion, and partitioning criteria). The demographics of the reference population should be representative of the patient population for which the RI will be used in making clinical decisions.

    1.1 Selection criteria must be defined. These are used to characterize the population and verify the health of individuals. They may include but are not limited to the following: (Walton 2001)
        - Biological (age, sex, breed, stage of reproductive cycle, production type)
        - Clinical (history and physical examination to establish health and husbandry practices)
        - Geographical (location) and seasonal (effects of temperature and day length) characteristics.

NOTE: Additional testing may be required to establish health. The type and extent of additional testing depends on the intended use of the proposed RI and may include, but are not limited to, a complete minimum database (CBC, biochemical profile, urinalysis), imaging, functional tests, fecal examination for parasites, lymphocyte phenotyping, and historical or clinical follow-up.

    1.2 Exclusion criteria are used to eliminate individuals that should not be included in the reference population. They may include but are not limited to the following:
        - Biological (e.g., fasted or non-fasted state, intense exercise, level of stress or excitement)
        - Physiological factors (illness, lactation, pregnancy, other). (Poole 1997) Evidence of illness within a defined period time preceding or following sample collection should be considered for exclusion of a reference individual.

- Administration of pharmacologically active agents. Individuals receiving pharmacologically active agents for treatment of specific disorders are usually excluded. However, routine administration of preventative dosages of anthelmintic medications typically is acceptable in most RI studies. (Poole 1997)

NOTE: Some of these inclusion and exclusion criteria may be used to partition reference populations into subgroups, for example, by age, sex, or reproductive status. (Walton 2001)

NOTE: Establishing health and exclusion criteria in wildlife species is particular challenging given the limited contact with these animals. Specific protocols for health assessment, restraint and sample collection based on prior experience with the species should be established and strictly followed to minimize unintended variation.

2. Develop a questionnaire that will establish whether a reference individual conforms to selection criteria, belongs to a partitioned subgroup, or should be excluded. The questionnaire is filled out by the owner and by the individual(s) examining the subject and collecting the sample. Owner consent for participation in the RI study is often included in this questionnaire and is mandatory in many institutions.

3. Consider the number of healthy reference subjects available to provide reference samples. Ideally a minimum of 120 reference individuals is available for determining nonparametric RI and 90% confidence intervals (CI) of the reference limits with enough extra individuals to allow for some rejection. Reference intervals determined from smaller sample sizes are commonplace and often necessary in veterinary medicine. However, thorough consideration for the effect of small sample size on the accuracy of population-based RI should be addressed early in the study. The smaller the sample size, the higher the degree of uncertainty in the estimation of the RI. Uncertainty is demonstrated by the width of the 90% CI around the upper and lower reference limits.

4. Reference individuals may be selected by either direct or indirect methods. Direct methods involve selection of known healthy individuals from a general population using specific criteria. Indirect sampling methods use medical databases containing results from both healthy and non-healthy individuals. Statistical and nonstatistical methods are employed to exclude samples from obviously unhealthy individuals, and RI are generated from the remaining values. Because RI established using indirect sampling methods inadvertently may include unhealthy individuals, RI derived in this manner may not accurately reflect the distribution of analyte quantities in a healthy population. In addition, information about preanalytical and analytical factors may not be available. Consequently, direct sampling methods are strongly recommended, and indirect sampling methods only should be used when other options for establishing RI are unavailable. Two types of direct sampling are possible:

4.1 *a priori* in which inclusion and exclusion criteria are established prior to selection of healthy reference individuals. This method is preferred when information about how biological and preanalytical factors affect the analyte(s) are well documented.

4.2 *a posteriori* in which inclusion and exclusion criteria are established after selection and testing of healthy reference individuals. This method typically is used when there is limited information about a new analyte or laboratory test, and it is not known how biological or preanalytical factors affect analyte quantities.

Preanalytical procedures – Patient preparation, sample collection and analytical quality assessment

5. Preparation of the reference individuals, sample collection, sample handling, and sample processing should be performed in a standardized manner that is consistent with the methods used for patient testing. In addition, consideration should be given to potential adverse effects of preanalytical factors in order to reduce variation that is not due to inter- or intra-individual variability. These details should be documented for future reference.

   5.1 Patient preparation and handling should be standardized (e.g. fasting or non-fasting, method of capture and restraint, use of sedatives or anesthesia).

   5.2 Sample collection (site, preparation of the site, vial type, collection system) and handling of the specimen (transportation, temperature, and centrifugation) should be standardized based on prior knowledge of the analyte. Sample type should be the same for all reference samples, e.g., all serum or all plasma.

   5.3 Sample collection may need to be standardized for time of day or for season, depending on the analyte being measured and the intended use of the RI. This is especially important for certain hormones. Alternatively, samples may need to be collected across seasons or throughout the day for a more general representation of analyte concentrations.

   5.4 Special sample handling requirements for certain analyte(s) must be known and strictly followed (e.g. on ice, anaerobic).

   5.5 Analyte stability should be determined prior to the RI study. This information is necessary to determine if certain analytes require specific storage conditions or analysis within a defined interval and if sample storage and batch analysis can be performed.

6. Estimates of analytical error (CV and bias) should be recorded for all methods. These may be determined during the RI study or during initial method validation (MV) studies. This information is necessary if transference of RI is utilized in the future. These estimates of analytical error should fall within the acceptable quality

requirement goals for imprecision (CV), inaccuracy (bias), and total allowable error (TEa) for existing methods or during method validation studies for new methods. Quality goals may be based on biologic variation, clinical interpretation of test results, consensus documents, or all of these. (Kenny 1999, Kjelgaard-Hansen 2010)

Analytical procedures

7. Analyze samples using methods that are monitored with strict quality control procedures. (Flatland 2010, ASVCP Quality Assurance Guidelines) Conditions for analysis should be well defined in a manner consistent with analysis of patient samples in order to reduce variation that is not due to inter- or intra-individual variability. However, variation that is part of everyday operation, such as changes in reagent lots and technical staff, should be integrated into RI studies whenever possible to approximate normal working conditions.

    7.1 Establish a laboratory submission policy for RI study samples.

    7.2 Establish rejection criteria for samples of inadequate quality.

    7.3 Monitor results in real time so that errors can be detected when re-measurement is still possible. This will prevent excessive rejection of reference values by reducing the number of potential outliers.

NOTE: Certain analytes in avian and reptiles may be unusual, and yet physiologic, during stages of the reproductive cycle, e.g., total calcium. Analytical interference with other analytes caused by these unusual concentrations must be known and accounted for during RI studies.

NOTE: Methods employed in establishing RI should be documented in detail, including the specific make and model of analyzer and the source of the reagents and quality control materials. This information should be retained with the RI study summary document for future reference.

Statistical analysis of reference values

8. Prepare and examine histograms of the reference values for initial assessment of distribution and identification of potential outliers.

NOTE: Boxplots are an alternative method of displaying data; however, they are not the preferred method for graphically presenting reference data.

9. Identify outliers. Outliers are values that do not truly belong to the underlying distribution of reference values. Inclusion of outliers will significantly affect reference limits; however, apparent outliers or unexpected values should not be eliminated indiscriminately.

9.1 Examine values that appear to be outliers in the histogram for errors. Errors may include, but are not limited to the following:

- Transcription errors
- Preanalytical and analytical errors, including those resulting from inclusion of inappropriate or poor quality samples (improper sample, hemolysis, lipemia)
- Inclusion of inappropriate reference individuals (animals subsequently proven to be unhealthy, animals that should have been excluded due to age, breed, physiology, etc.)

NOTE: Values resulting from these types of errors should be eliminated whether or not they are located within the extremities of the distribution.

9.2 Use an appropriate statistical method to further examine the reference data for outliers. The following are 2 of the most commonly used tests to detect outliers in reference data, although other methods are available (Grubb 1950, Jain 2010). Optimal performance of these tests requires that the data approximate a Gaussian distribution. (Horn 2001) Therefore, application of these tests typically occurs after data has undergone transformation (if not Gaussian) and is tested for normality.

- Dixon's outlier range statistic typically identifies the single most extreme value at the upper or lower limit as an outlier. (Dixon 1983) The simplest criterion of rejection (r criteria) is $D/R > 0.3$, where D is the absolute difference between the most extreme value and the next nearest value divided by the range of all values (R) including the extreme value(s). (Reed 1971) This criterion is fairly conservative and favors retention of reference values. If several values at one extremity appear to be outliers, the least most extreme value can be treated as the most extreme for calculating the ratio (block procedure). If the least most extreme value is determined to be an outlier, all the more extreme values can be eliminated. (Barnett 1978) The ratio also can be compared to published tables of critical values for more stringent evaluation of outlier status. (Rorabacher 1991) Critical values vary depending on the number of anticipated outliers at one or both extremities (one- or two-tailed), the number of reference values, and the desired level of confidence in detecting outliers.
- Horn's algorithm using Tukey's interquartile fences identifies multiple outliers located at the upper and lower extremities. (Horn and Pesce 2003, Horn and Pesce 2005) The criterion for rejection is values exceeding interquartile (IQ) fences set at Q_1 - 1.5*IQR and Q_3 + 1.5*IQR (IQR = interquartile range; $IQR = IQ_3 - IQ_1$ where IQ_1 and IQ_3 are the 25^{th} and 75^{th} percentiles, respectively). This test is more stringent than Dixon's range statistic, which favors retention.

9.3 Correct known errors in outliers if possible.

9.4 Eliminate outliers proven not to belong to the reference population by statistical or other means.

9.5 Once outliers are eliminated, retest the remaining data for additional outliers.

NOTE: The RI study summary document should reflect the number of outliers eliminated, the method used to identify them and the final number (n) from which RI are determined.

9.6 When data do not approximate Gaussian distribution or cannot be transformed to Gaussian distribution, these procedures do not perform optimally and may erroneously identify values located in the tail of skewed distributions as outliers, such as occurs with enzyme activities. (Horn 2001) When data are non-Gaussian, nonparametric methods should be used to establish RI (see below). Because nonparametric methods establish reference limits by trimming the most extreme values, outliers have less of an effect on the RI than with parametric and robust methods. (Horowitz 2008)

Not all outliers and inappropriate values will be detected with these statistical methods. (Solberg & Lahti 2005) Multiple outliers located at one or both extremities may have the effect of masking the presence of outliers and rendering these methods unsuitable for outlier detection. (Horn 2001) The best means by which to avoid inclusion of inappropriate values within the reference data is to ensure that all reference individuals are healthy and belong to the desired demographic and to avoid unintended preanalytical and analytical variation. The challenge of correctly identifying outliers is magnified in wildlife RI studies where health is difficult to substantiate and where manual or field methods may introduce a higher level of imprecision and inaccuracy.

When health is well-defined and can reasonably be substantiated, outlier tests that favor retention of reference values are preferred (e.g., Horn's algorithm using Tukey's IQ fences and Dixon's range statistic with confidence levels of $\alpha = 0.05$). This is analogous to low Type I error. However, when "convenience samples" are used or when health is difficult to substantiate, as is often the case with wildlife RI studies, an outlier test that more readily identifies a value as outlier should be applied in order to exclude potentially erroneous values (low Type II error). This would include comparing Dixon's r criteria to a table of critical values with confidence levels of $\alpha = 0.1$. (Rorabacher 1991)

In conclusion, attempts to identify and eliminate outliers should be made during the analysis of reference values. Extreme values should not be eliminated indiscriminately because these values may represent the true distribution of values in a healthy population. However, inclusion of true outliers can significantly affect the determination of reference limits rendering the RI less representative of the desired demographic. Clinical experience also must be employed in determining when to retain or eliminate values at the extremities of the distribution.

10. If parametric or robust methods will be used to determine RI, use a goodness-of-fit test to determine if the distribution of the reference data is Gaussian. One of the following methods may be selected:
    - Anderson-Darling
    - Kolomogorov-Smirnov
    - Shapiro-Wilk
    - D'Agostino and Pearson omnibus

    If distribution is not Gaussian, transform the data using an appropriate function (e.g., log or Box-Cox transformation) and reassess the distribution. If Gaussianity cannot be established after several transformation trials, parametric methods cannot be used to establish RI. Although best applied to data with a symmetrical distribution, the robust method may be used when reference values do not exhibit Gaussian distribution. (Horowitz, 2008; Horn & Pesce, 2005, Chapter 6, page 47-57)

11. Select statistical method for analysis based on the number and distribution of reference values.

    11.1 Nonparametric methods do not require assumption of a particular distribution of the data and are recommended when ≥120 reference samples are available. Nonparametric methods typically encompass the central 95th percentile of reference values and use the 2.5th and 97.5th fractal as the lower and upper reference limit, respectively. (Horn and Pesce 2003) Nonparametric determination of 90% confidence intervals (CI) of the lower and upper reference limits is possible with ≥120 reference samples. Although robust methods are preferred when there are <120 reference values (see below), nonparametric methods can be used when the distribution of reference data is not Gaussian; however, 90% CI cannot be calculated nonparametrically and alternative methods of determining CI must be used, e.g., bootstrap. (Horowitz 2008) The minimum number of reference values required to determine central 95% RI nonparametrically is 39.However, in this situation, the most extreme values serve as the lower and upper reference limits. (Horn and Pesce 2005) If a method to detect outliers is not performed, sufficient numbers of samples should be collected to allow trimming of ≥ 2 values at both extremities to avoid inclusion of potential outliers.

    11.2 Robust methods are recommended when $40 \leq x \leq 120$ reference samples are available. The robust method utilizes an iterative process to estimate location and spread of the data. (Horn & Pesce 2005, 1999, 1998) Although the robust method performs best when reference data has a symmetrical distribution (with or without transformation), it can be used in the absence of Gaussianity. Ninety percent CI around the reference limits should be determined using bootstrap methods. The robust method is included in several clinical laboratory software programs, including CBstat (CBStat), Reference Value Advisor freeware (Reference Value Advisor, Geffre et al 2011), and MedCalc (MedCalc).

11.3 Parametric methods may be used when $40 \leq x \leq 120$ reference samples are available and the data has Gaussian distribution or can be transformed to Gaussian distribution. Parametric methods encompass slightly more than the central 95percentile of the data and establish the lower and upper reference limits at mean minus 2SD and mean plus 2SD, respectively (SD = standard deviation). Ninety percent CI around the reference limits should be determined using parametric methods.

11.4 There will be instances in veterinary medicine when a limited number of reference samples can be collected. Examples include neonates, special species, zoological species and wildlife. When $20 \leq x < 40$ reference samples are available, RI should be calculated by robust (distribution independent) or parametric (if Gaussianity can be established) methods. To highlight the uncertainty in the upper and lower reference limits resulting from small sample sizes, 90% CI should be calculated. In addition, the following should be available to allow informed clinical decisions to be made:

- A histogram of the data
- Mean (if Gaussian) or median (if not Gaussian)
- The minimum and maximum values or a table of all reference values listed in ascending order.

11.5 Reference intervals should not be determined when < 20 reference samples are available. When $10 \leq x \leq 20$ reference samples are available, the following information should be reported to aid in clinical decision making.

- A table of all reference values listed in ascending order
- A histogram of the data for graphic visualization
- Mean (if Gaussian) or median (if not Gaussian)

Reference values from < 10 individuals should not be reported because sample sizes this small are unlikely to be representative of the distribution of a variable within a population. When <10 reference individuals can be collected, subject-based reference intervals should be considered (see the section on Biological variation, individuality, and subject-based reference intervals).

NOTE: Although not ideal, it is acknowledged that situations occur in veterinary medicine in which < 40 reference samples are available. When this occurs emphasis should be placed on collecting samples that are free from unintended variability by paying strict attention to selection of suitable reference subjects and adherence to standardized collection techniques and well controlled methods of analysis. Evaluation for the presence of outliers is particularly important with small sample sizes, because the presence of a single outlier has a significant effect on estimated reference limits. If outliers are eliminated, every attempt should be made to collect replacement samples.

NOTE: Reference intervals calculated by several of the methods listed above can be compared. If the RI are similar, any of the statistical methods is acceptable. (Horn 1998) When RI differ, clinical judgment may be required to select the most appropriate

reference limits with some experts recommending selection of the narrower RI in order to limit the number of false negative results. (Horn 1998)

NOTE: Confidence intervals around the upper and lower reference limits should be calculated whenever sample size permits. Confidence intervals provide an estimate of the uncertainty of the reference limits and are generally narrower for larger sample sizes than for smaller sample sizes. Boyd and Harris recommend that CI should not exceed 0.2 times the width of the RI. (Harris 1995, Horowitz 2008) When CI exceed this limit, an effort to collect additional reference samples should be made.

12. Determine the need for partitioning into subclasses based on physiological differences that are expected to result in important clinical differences in RI. Partitioning favors homogeneous sub-populations, decreasing variability between individuals and narrowing the RI. However, partitioning only should be considered if there is a minimum of 40 individuals within each subclass or if there are clear clinical reasons. Partitioning criteria should consider not only the subgroup means, but also subgroup standard deviations (SD). More recent partitioning criteria examine the proportion of each subgroup that fall outside the upper and lower limits of a combined RI (Lahti 2004, Ceriotti 2009). This proposal is based on the fact that RI generally encompass 95% of reference values and exclude 2.5% of values at the upper and lower extremities. Consideration also must be given to unequal sizes of the subgroups within the general population as well as within the reference sample group. (Lahti 2002 – partitioning) The following are some simple recommendations:

    12.1 Partitioning is recommended if the absolute difference between subgroup means exceeds 25% of the reference range (range = upper limit – lower limit) of the combined central 95% RI. (Sinton, 1986) This method is conservative in recommending partitioning and requires Gaussian distribution of data for optimal performance. (Lahti 2004 existing methods)

    12.2 Partitioning is recommended if the ratio of subgroup SD (larger SD/small SD) exceeds 1.5, regardless of the subgroup means. (Harris and Boyd 1990) If this criterion is exceeded, Harris and Boyd recommend that subgroup means be compared by the standard normal deviate test. If there are 120 samples in each subgroup, partitioning is recommended if $z > 3$ (critical z-statistic). The z value is calculated as follows:

    $$z = \text{mean}_1 - \text{mean}_2 \,/\, [(SD_1^2/n_1) + (SD_2^2/n_2)]^{1/2}$$

    If there are fewer than 120 samples in each subgroup, z should be compared to an alternative critical z-statistic based on the average size of the subgroups (alternative critical *z*-statistic = $3 \times (n_{average}/120)^{1/2}$). This procedure works optimally when the data has a Gaussian distribution and the subclasses are of similar size and SD. (Lahti 2004 – existing)

    12.3 Partitioning recommendations by Lahti et al for Gaussian distributions begin with the same subgroup SD ratio criterion of >1.5. However, if the SD ratio is ≤ 1.5, then the differences between the upper (D_U) and lower limits (D_L) of the

2 subgroups should be examined in relation to the smaller of the subgroup SD ($SD_{smaller}$). (Lahti 2002 – Gaussian)

$$D_U = URL_1 - URL_2 \text{ (use absolute values)}$$
$$D_L = LRL_1 - LRL_2$$

Partitioning is not recommended when both D_L and $D_U < 0.25\ SD_{smaller}$
Partitioning is recommended when either D_L, D_U or both $\geq 0.75\ SD_{smaller}$
When either D_U or D_L or both fall in between these criteria, the decision on whether to partition is made using non-statistical criteria.

12.4 The partitioning recommendations above, while reasonably simple to calculate, contain several weaknesses. (Lahti 2004 – existing methods) To correct for these weaknesses, Lahti et al made the following general recommendations for partitioning; however, application of these criteria is more challenging. (Lahti 2004) Partitioning is recommended if >4.1% or <0.9% of a subgroup falls outside the upper or lower limits of a combined RI. If 1.8% < x < 3.2 % of a subgroup falls outside the combine RI, partitioning is not recommended. When the proportions of a subgroup outside a combined RI are between these criteria, the decision to partition the RI is made using non-statistical criteria.

12.5 Non-statistical criteria that support partitioning include:

- If descriptors used to assign a patient to a partitioned subgroup are easily obtainable and maintained in the patient record.
- If reference limits serve as critical clinical decision limits.
- If the literature documents important clinical differences between subgroups

13. Document all previous steps and procedures so that RI are clearly defined. A complete and detailed RI study summary document should be available to users upon request (see Table 3). The laboratory should retain RI summary documents for a pre-determined length of time or indefinitely. These details also should be included in publications of RI in veterinary species to allow critical evaluation by potential users of the reference data. Addendum 1 contains tables for hematology and biochemistry that can be used to report reference value data for published studies.

14. Reference intervals determined de novo within a laboratory should be reviewed every 3 to 5 years and re-validated if needed. Re-validation is necessary anytime there has been a change in assay methodology or a significant change in patient populations. (See the section on transference and validation.)

Postanalytical procedures – Laboratory presentation of reference intervals

15. Information included in a laboratory report should aid the clinician in the medical decision making process and be presented in a clear and concise manner.

15.1 Reference intervals typically are printed on the patient report; however, this should occur only when the RI is applicable to that patient.

15.2 Reference intervals that deviate from customary percentiles and limits or are specific to a certain subclass (age, sex) should clearly be identified on the report.

15.3 It is useful to indicate which patient values are increased or decreased in comparison to the RI.

15.4 Information that may be important for clinical decision-making, but that cannot be contained within the patient's laboratory report, should be available to the clinician in a written report (RI study summary document). This information may include, but is not limited to: (Plebani)

- Reference population demographics and number of reference subjects sampled
- Subject preparation and time or season of collection, if relevant
- Sample type and handling
- Confidence intervals around the reference limits

15.5 Changes in RI owing to introduction of new methods or analyzers or adjustments to RI resulting from changes in population demographics should be communicated to all users.

TRANSFERENCE AND VALIDATION OF REFERENCE INTERVALS

In order to forego the expense and difficulty of establishing intra-laboratory RI, many laboratories adopt RI from other laboratories or from instrument manufacturers. The following procedures are recommended for the transference and validation of RI adopted from other sources.

1. Several issues should be scrutinized when external RI are considered for transference.

1.1 The appropriateness of the reference population with respect to age, sex, breed, geography, physiology, etc.

1.2 Differences in pre-analytical techniques, such as patient preparation and collection method.

1.3 Differences in test methodology

1.4 Differences in instrument accuracy and imprecision (analytical quality).

1.5 Differences in laboratory quality by the laboratory donating the RI and the laboratory adopting them.

If significant differences are detected in the above areas, transference may not be appropriate. Many RI studies are not well documented and often lack the detailed information necessary to determine the appropriateness of transference.

2. Estimates of analytical error (CV and bias) are necessary to determine transferability. RI can be transferred between laboratories using different analytical methods as long as the methods possess similar analytical quality (see section 3.5). In addition, a transferred RI only remains valid as long as the laboratory maintains the original precision (CV) and accuracy (bias) that were used to establish transference validity. (Ceriotti 2009)

3. A comparison of methods study may be used to determine if analytical methods are similar. (Jensen 2006 and ASVCP QC Guidelines website)

    3.1 Analytical methods are comparable if the slope of the regression line generated during the comparison of method study approximates 1.0 and the y-intercept is small relative to the data range and RI. If comparable, RI can be transferred directly.

    3.2 If systematic difference (bias) exists between analytical methods, regression statistics can be used to adjust the upper and lower reference limits. This commonly is used when a laboratory changes instruments or methods and transfers their existing RI to the new instrument or method. Transference using regression statistics only should be employed for one change of instrumentation or methodology. Depending on the distribution of data, alternate statistics may be required, e.g., Passing-Bablock or Deming regression or difference of means for analytes with narrow distributions, such as electrolytes.

4. Once it is determined that a RI is suitable for transference, validation of the transferred RI may be accomplished by one of the following procedures:

    4.1 Evaluating 20 samples representative of the laboratory's own patient population against the candidate RI. (Horowitz 2008) This validation procedure is relatively quick and straightforward. These 20 values first should be examined for outliers. If outliers are detected, they should be eliminated and additional samples collected. If ≤ 2 of the 20 values fall outside the candidate RI, it is considered transferable. If 3 or 4 values fall outside the RI, another 20 patients can be tested and interpreted in the same manner as the original 20 samples. If > 4 of the original 20 values fall outside the candidate RI, transference is rejected for that analyte. This is basically a binomial test and will not determine whether the transferred RI is too wide for the receiving laboratory. If all 20 samples fall within the candidate RI, it may be inappropriately wide for the adopting laboratory. When this occurs, another 20 samples should be evaluated. The probability that all 20 results will be within the candidate RI is about 0.36, and with 40

samples the probability falls to about 0.13. If all 40 samples fall within the candidate RI, then it is likely too wide and de novo RI should be determined. (Horn and Pesce 2005, Chapter 10, page 77-80.)

4.2 A more thorough validation uses results from 40-60 healthy subjects from the laboratory's patient population. If individual results are available from the original RI study, the reference values from both groups can be compared using more sophisticated statistical equations (Mann-Whitney U test, median test, Siegel-Tukey test, Kolmogorov Smirnov test). (Horowitz) With the introduction of the robust method, acceptable RI can be generated from 40-60 reference samples, obviating the need for transference and validation using this method.

4.3 Subjective assessment of the quality and applicability of the RI. This requires complete and detailed documentation of the original RI study to ensure that all procedures used in the donating laboratory are equivalent or comparable to the adopting laboratory. Because complete and detailed information is infrequently available, this validation procedure is seldom sufficient.

NOTE: The clinical use of the test should be considered when selecting the method for validating a transferred RI, e.g., for more critical tests, procedure 4.2 or 4.3 should be used.

5. It is the responsibility of the end user to validate RI provided by manufacturers of veterinary diagnostic tests and analyzers. Manufacturers may generate RI using a single analyzer or multiple analyzers at one or more locations (see multicenter RI below). Manufacturers should provide sufficient information regarding the RI so that their clients can assess the quality and applicability of the RI study to their patients. Information provided by the manufacture should include, but is not limited to the following (see also Table 3):
   - Selection method of reference subjects (direct or indirect)
   - Demographics of the reference sample group
   - Preanalytical and analytical factors
   - Number of reference values used to establish RI
   - Statistical methods used to identify outliers and generate RI

6. Indications for re-validation of RI include but are not limited to:
   - Excessive false positive and false negative results
   - Significant changes in patient populations, preanalytical techniques, or analytical quality
   - Periodic reassessment of RI every 3-5 years

COMMON (or multicenter) REFERENCE INTERVALS

Another alternative to determining RI de novo within each laboratory is for several laboratories serving a similar patient population to contribute to the generation of common (or multicenter) RI. The following summarizes the advantages and important considerations of this approach. (Petersen 2004, Horowitz 2008, Jones 2004)

1. Advantages of common RI include:

    1.1 Standardization of results, reporting, and interpretation in a mobile population in which veterinary patients may change locations and laboratories during their lifetimes.

    1.2 Ability to recruit large numbers of reference samples from multiple participating laboratories.

    1.3 Large numbers of reference values allow partitioning according to desired criteria (age, sex, breed, use, activity, other).

    1.4 Distribute cost of establishing RI among multiple laboratories.

2. Laboratories participating in a common RI study should adhere to the following procedures:

    2.1 Although is it preferred, laboratories contributing results to a common RI do not have to have the same analyzer (manufacturer and model number). However, all analyzers must be calibrated to produce comparable results (Jensen 2006) and must meet the same quality requirements. Common calibration and performance quality are essential if the resulting RI are to be used by all participating laboratories. If common RI are adopted by a laboratory that did not participate in the study, the common RI must be validated, even if the laboratory uses the same analyzer as was used in the study.

    2.2 Determine if similar populations are present among the participating laboratories – a prerequisite for the use of common RI.

    2.3 Establish uniform selection and exclusion criteria for defining reference individuals and establishing health, as detailed in section 1 of the guidelines for establishing de novo RI.

    2.4 Standardize pre-analytical and analytical factors, as detailed in sections 5 and 7 of the guidelines for establishing de novo RI.

2.5 Establish quality requirement goals for imprecision (CV) and inaccuracy (bias) for all analytes. These may be based on biologic variation, clinical interpretation of test results, or both. (Kenny 1999, Kjelgaard-Hansen 2010)

2.6 Bias, in particular, must be controlled and minimized by each laboratory. If bias varies significantly among participating laboratories, the use of common RI may lead to clinical misclassifications. A 'maximum allowable bias' should be established based on calculated TE (TEc = bias + 2CV) using the maximum CV obtained by participating laboratories. If a laboratory exceeds this predetermined bias limit, measures should be initiated (re-calibration, etc.) to return bias to limits that allow continued use of the common RI.

2.7 Use calibrators traceable to an international standard to determine the trueness and comparability of measurements across all participating laboratories. Alternatively, a single, pooled specimen may be used as a common calibrator.

2.8 Once common RI are established, variability caused by changes in calibrator lots must be minimized. Instead of using the mean assigned to the calibrator by the manufacturer, a mean value for a common calibrator should be determined from results submitted by all participating laboratories. Each laboratory establishes a mean calibrator value by repeat analysis (n = 10 to 20 repetitions) and then submits this mean for determination of a common mean.

2.9 Validate quality control procedures designed for high probability of error detection and low probability of false rejection in order to monitor and maintain stable performance.

2.10 If possible, use the same quality control materials and reagents to facilitate standardization and comparison of results. For analyzers with unique QC materials (hematology analyzers), this may not be possible.

NOTE: Manufacturers of diagnostic equipment may provide users with 'common RI' determined from one or more analyzers; however, RI validation is recommended to ensure that local patient populations are represented by the common RI and that performance of the on-site analyzer meets calibration and quality performance requirements established in the common RI study.

USE OF PUBLISHED REFERENCE INTERVALS

Interpreting clinical data using inappropriate RI may lead to misclassification of a patient, which can result in misdiagnoses, improper treatments or both. Unless a RI is representative of the patient's demographics and is determined using similar preanalytical procedures and comparable analytical methods, it is not appropriate as a diagnostic reference for clinical decision-making. Reference intervals published in textbooks,

journal articles, and web-based databases or provided by instrument manufacturers may or may not contain sufficient information to determine whether the RI is appropriate for a particular patient. In addition, the quality of published RI is quite variable. In addition to information regarding preanalytical and analytical procedures, the quality of a RI depends on the reference sample size, the use of procedures to detect and eliminate outliers, and correct use of statistical procedures to estimate the reference limits. Published RI should be used with caution, and only when sufficient information is available to determine their quality and applicability to the patient and analytical method. If published RI are adopted for extended use, appropriate validation procedures should be performed as described in this document.

BIOLOGICAL VARIATION, INDIVIDUALITY AND SUBJECT-BASED REFERENCE INTERVALS

Population-based reference intervals serve as a comparison for patient test results when an alternative frame of reference is not available. However, due to relatively high inter-individual variability, population-based reference intervals sometimes lack necessary sensitivity to detect changes in the health of an individual. The following recommendations provide guidance for the appropriate use of subject-based RI.

1. Subject-based RI are more sensitive than population-based RI for disease detection when intra-individual biologic variation, or coefficient of variation of an individual (CV_I), is less than inter-individual variation, or CV of a group (CV_G). (Fraser 2004)

    1.1 The index of individuality provides an objective criterion to determine the relative utility of subject-based versus population-based RI. A quantitative measure of individuality, the index of individuality is represented by the equation $(CV_I^2 + CV_A^2)^{1/2} / CV_G$, where CV_A is analytical variation (random error or imprecision). Because $CV_A < CV_I$ for many automated assays, the index of individuality is often simplified as CV_I/CV_G. (Fraser 2004)

    1.2 When the index of individuality is < 0.6, a subject-based RI is preferable to a population-based RI, whereas when it is >1.4, individual RI yield no more information than population-based RI. (Fraser and Harris 1989) For patient monitoring, using an index < 0.48 rather than <0.6 increases the probability of the subject-based RI detecting change in a monitoring situation. (Iglesias Canadell et al. 2005)

    1.3 With subject-based RI, a reference change value (RCV) serves to determine whether a difference between consecutive measurements in an individual is significant ($p \leq 0.05$). Reference value change also is called the critical difference. (Jensen 1993) The RCV is based upon CV_I values in health and the dispersion of these variations across a population: $RCV = z \times [2(CV_I^2 + CV_A^2)]^{1/2}$ where z represents the z-statistic. RCV is best applied when $CV_A < 0.5 \times CV_I$. Because $CV_A << CV_I$ for many automated assays, RCV simplifies as $z \times 2^{½} \times CV_I$ or $z \times (1.41CV_I)$.

NOTE: Laboratory values determined during optimal health are required to serve as a base-line for identifying RVC that may indicate an illness. Alternatively, RVC between serial laboratory results can be used to monitor progression or resolution of a disease.

1.4 The *z*-statistic conventionally used for RCV calculation is $z = 1.96$, which provides a 50% probability of detecting an increase with a 5% probability of type I error. When larger *z*-factors are used, such as $z = 3.34$, the probability of a significant change being detected increases to 90% while increasing the probability of Type II error (Iglesias Canadell et al. 2004)

2. To use subject-based RI, it is necessary to know the CV_I for each analyte, as well as the imprecision of the analytical method for that analyte.

2.1 Published CV_I are available for many analytes in the dog (Jensen and Kjelgaard-Hansen, Wiinberg) and for some exotic species (Bertelsen).

2.2 Biologic variation can be measured with only a small number of healthy reference individuals when care is taken to minimize pre-analytical variation; thus, these data may readily be measured by a single laboratory when published biologic variation data are not available. (Fraser and Harris)

2.3 Use of CV_I and RVC derived from healthy individuals to determine significant changes in analyte quantities in patients with stable chronic diseases may not be appropriate. Consequently, estimates of CV_I from patients with chronic stable disease are being developed in human medicine to better monitor chronic disease conditions. (Ricos 2007) This information currently is not available in veterinary medicine.

ESTABLISHING DECISION THRESHOLDS (or decision limits)

The clinical utility of a test depends partly on diagnostic accuracy – that is, the ability of the test to discriminate between patients with disease and without disease. Diagnostic accuracy, in turn, depends on the decision limit used for determining a positive or negative result. Unlike RI, which are "defined by statistical methods and are descriptive of specific populations, decision thresholds are defined by consensus and distinguish among different populations". (Horowitz 2008) The following procedure outlines the steps and basic principles required for designing prospective studies to establish diagnostic thresholds. (Jorgensen 2004, Peblani 2004, Zwieg 1995)

1. Define the clinical question with regards to the following factors:

1.1 Characterize the target population as to the frequency and duration of disease, as well as age, sex, and other relevant information that aids in interpretation of test results.

1.2 State the management decision to be made concerning the disease in question.

1.3 Identify the role of the test in the clinical decision-making process with regards to the disease.

2. Prospectively select individuals (study sample) that are representative of the relevant target population. The study sample should consist of subjects that are anticipated to test both positive and negative for the test under investigation. Consult a statistician to establish the size of the study sample required for statistically relevant results.

2.1 Diversity within the target population should accurately represent what is expected in clinical practice.

2.2 If the clinical question concerns the presence or absence of a disease, a statistically relevant number of individuals should have the disease in question.

2.3 If the clinical question concerns determining the severity or prognosis of a disease, the entire sample may consist of diseased animals representing the entire spectrum of disease severity or outcome.

2.4 To prevent bias, selection of study subjects should be independent of test results for the test or analyte being evaluated.

3. Whenever possible, study subjects should be tested with the test under investigation without knowledge of their disease classification to prevent prejudiced decision-making.

3.1 When comparing performance of multiple tests, tests should be executed on all subjects at the same time or at the same stage of disease, and all tests should be performed on the same sample.

3.2 If sample stability permits, assaying samples in a single batch minimizes between-run analytical variance.

3.3 Subjects with unexpected test result should not be excluded from the study to avoid bias favoring test performance.

4. Comprehensive examination and alternative testing are used to establish true disease positive and true disease negative status. For tests evaluating disease severity or prognosis, standardized staging, grading, or scoring schemes (Hayes, 2010) and common outcome assessments should be used.

5. Receiver-operating characteristic (ROC) curve are created by plotting the sensitivity and specificity of the test under investigation at a variety of decision thresholds.

(Gardner 2006, Stephan 2003) ROC curves are used to compare performance to select decision thresholds that optimally answer the clinical question or satisfy the diagnostic requirement.

5.1 The area under the curve (AUC) is an estimate of test accuracy. Under most circumstances when comparing multiple tests, the more accurate test has the higher AUC.

5.2 Optimal decision limits are selected by location on the ROC plot (most upper left point) or by determining the decision limit with the highest proportion of correct interpretations (most true positive and true negative results).

5.3 Confidence intervals around points on a ROC curve can be derived parametrically and non-parametrically.

6. Further confirmation of the true clinical state of each subject should be done during or after the completion of the study without utilizing the results of the test under investigation.

6.1 This provides quality control crosscheck to insure that previous criteria for disease classification are accurate.

6.2 Confirmation procedures may include histopathology (surgical biopsy or necropsy) or preferably long-term follow-up regarding the course of disease.

Summary and closing

A uniform and consistent approach to establishing RI will benefit the entire veterinary medical community. The acceptance of statistical methods for establishing RI from small samples sizes, the approval of transference and validation as an accepted method for obtaining RI, and a growing interest in common RI will expand the ability of veterinary diagnostic laboratories to provide RI for a variety of species and distinct subgroups. Adherence to these guidelines by all those establishing and publishing RI in animals should facilitate communication within the broader veterinary community. By providing detailed information in RI study summaries, judicious use of published RI may be possible. In the absence of appropriate, population-based RI, subject-based RI may be a viable alternative for interpreting clinical data in certain situations.

References

ASVCP Quality Control Guidelines http://www.asvcp.org/pubs/pdf/ASVCPQualityControlGuidelines.pdf Accessed September 22, 2011

Barnett V, Lewis T. *Outliers in Statistical Data*. Chichester, England: John Wiley and Sons, Ltd.; 1978;381-386.

Bertelsen MF, Kjelgaard-Hansen M, Howell JR, Crawshaw GJ. Short-term biological variation of clinical chemical values in Dumeril's monitors (*Varanus dumerili*). J Zoo Wildl Med 2007;38(2):217-21.

Ceriotti F, Hinzmann R, Panteghini M. Reference intervals: the way forward. Ann Clin Biochem 2009;46(Pt 1): 8-17.

CBstat http://direct.aacc.org/ProductCatalog/Product.aspx?ID=2475 Accessed September 22, 2011.

Dixon WJ. Processing data for outliers. Biometrics 1983;9:74-89. (corrections to this article were made by Rorabacher)

Dybkær R. Approved recommendation (1987) on the theory of reference values. Part 6. Presentation of observed values related to reference values. Clin Chim Acta 1987;170:S33-S42.

Flatland B, Freeman KP, Friedrichs KR, Vap LM, Getzy KM, Evans E, Harr KE. ASVCP quality assurance guidelines: control of general analytical factors in veterinary laboratories. VCP 2010;39:264-277.

Fraser CG. Inherent biological variation and reference values. Clin Chem Lab Med 2004;42:758-764.

Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. Crit Rev Clin Lab Sci 1989;27:409-437.

Friendship RM, Lumsden JH, McMillan I, Wilson MR. Hematology and biochemistry reference values for Ontario Swine. Can J Comp Med 1984;48:390-393.

Gardner IA, Greiner M. Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. Vet Clin Pathol 2006;35:8-17.

Geffre A, Concordet D, Braun JP, Trumel C. Reference Value Advisor: a new freeware set of macroinstructions to calculate reference intervals with Microsoft Excel. Vet Clin Pathol. 2011;40:107-112.

Geffré A, Friedrichs K, Harr K, Concordet D, Trumel C, Braun JP. Reference values: a review. Vet Clin Pathol 2009;38(3):288-98.

Gräsbeck R, Saris NE. Establishment and use of normal values. Scand J Clin Lab Invest. 1969;26(Suppl 110):62-63.

Grubb FE. Sample criteria for testing outlying observations. Annal Math Stat 1950;21:27-58.

Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranged. Clin Chem 1990;36(2):265-270.

Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker, 1995:77-91.

Hayes G, Mathews K, Kruth S, Doig G, Dewey C. Illness severity scores in veterinary medicine: what can we learn? J Vet Intern Med 2010;24:457-466.

Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and nonhealthy individuals on reference interval estimation. Clin Chem 2001;47(12):2137-2145.
Horn PS, Pesce AJ. Reference intervals: an update. Clin Chim Acta 2003;334:5-23.
Horn PS, Pesce AJ. Reference intervals: a user's guide, AACC Press, Washington DC, 2005.
Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem 1998;44(3):622-631.
Horowitz GL, Altaie S, Boyd J, Ceriotti F, Gard U, Horn P, Pesce A, Sine H, Zakowski J. Clinical and Laboratory Standards Institute (CLSI). Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guidelines, 3rd ed, CLSI document C28-A3, Vol. 28, No. 3, 2008.
Iglesias Canadell N, Hyltoft Petersen P, Jensen E, Ricos C, Jorgensen PE. Reference change values and power functions. Clin Chem Lab Med 2004;42:415-422.
Iglesias Canadell N, Hyltoft Petersen P, Ricos C. Power function of the reference change value in relation to cut-off points, reference intervals and index of individuality. Clin Chem Lab Med 2005;43:441-448.
Jain RB. A recursive version of Grubb's test for detecting multiple outliers in environmental and chemical data. Clin Biochem 2010;43:1030-1033.
Jensen AL, Aaes H. Critical differences of clinical chemistry parameters in blood from dogs. Res Vet Sci 1993;54:10-14.
Jensen AL, Kjelgaard-Hansen M. Method comparison in the clinical laboratory. Vet Clin Pathol 2006;35:276-286.
Jones GRD, Barker A, Tate J, Lim C, Robertson K. The case for common reference intervals. Clin Biochem Rev 2004;25:99-104.
Jorgensen LG, Brandslund I, Hyltoft Petersen P Should we maintain the 95 percent reference intervals in the era of wellness testing? A concept paper. Clin Chem Lab Med 2004;42:747-51.
Kenny D, Fraser CG, Hyltoft Petersen P, Kallner A. Consensus agreement. Scand J Clin Invest 1999;59:585.
Kjelgaard-Hansen M, Jensen AL. Subjectivity in defining quality specifications for quality control and test validation. Vet Clin Pathol 2010;39:133-135.**Lahti A.** Partitioning biochemical reference data into subgroups: comparison of existing methods. Clin Chem Lab Med 2004;42:725-733.
Lahti A, Hyltoft Petersen P, Boyd JC. Impact of subgroup prevalences on partitioning Gaussian-distributed reference values. Clin Chem 2002;48:1987-99.
Lahti A, Hyltoft Petersen P, Boyd JC, Fraser C, and Jørgensen N. Objective criteria for partitioning Gaussian distributed reference values into subgroups. Clin Chem 2002;48(2):338-352.
Lahti A, Hyltoft Petersen P, Boyd JC, Rustad P, Laake P, Solberg HE. Partitioning of nongaussian-distributed biochemical reference data into subgroups. Clin Chem 2004;50(5):891-900.
Lumsden JH, Mullen K. On establishing reference values. Can J Comp Med 1978:42:293-301.
Lumsden JH, Mullen K, McSherry BJ. Canine hematology and biochemistry reference values. Can J Comp Med 1979;43:125-131.

Lumsden JH, Mullen K, Rowe R. Hematology and biochemistry reference values for female Holstein cattle. Can J Comp Med 1980;44:24-21.
Lumsden JH, Rowe R, Mullen K. Hematology and biochemistry reference values for the light horse. Can J Comp Med 1980;44:32-42.
MedCalc http://www.medcalc.org/index.php Accessed September 22, 2011.
Petersen H, Rustad P. Prerequisites for establishing common reference intervals. Scand J Clin Lab Invest; 2004; 6; 64(4):285-92.
PetitClerc C, Solberg HE. Approved recommendation (1987) on the theory of reference values. Part 2. Selection of individuals for the production of reference values. Clinica Chimica Acta; 1987; 170(2-3):S1-S11.
Plebani M. What information on quality specifications should be communicated to clinicians, and how? Clin Chim Acta 2004;346(1):25-35.
Poole T. Happy animals make good science. Lab Anim; 1997; 31(2):116-24.
Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate or normal range. Clin Chem 1971;17:275-284.
Reference Value Advisor http://www.biostat.envt.fr/spip/spip.php?article63 Accessed September 22, 2011.
Ricos C, Iglesias N, Garcia-Lario J, et al. Within-subject biological variation in disease: collated data and clinical consequences. Ann Clin Biochem 2007;44:343-352.
Rorabacher DB. Statistical treatment for rejection of deviant values: critical values of Dixon's Q parameter and related subrange ratios at the 95% confidence level. Anal Chem 1991;63(2):139-146.
Sinton TJ, Cowley D, Bryant SJ. Reference intervals for calcium, phosphate, and alkaline phosphatase as derived on the basis of multichannel-analyzer profiles. Clin Chem 1986;32:76-79.
Solberg HE. Approved recommendation (1986) on the theory of reference values. Part 1. The concept of reference values. Clin Chim Acta 1987;165:111-118.
Solberg HE. Approved recommendation (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. Clin Chim Acta 1987;170:S13-S32.
Solberg HE. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation of individuals and collection of specimens for the production of reference values. Clin Chim Acta 1988;177:S1-S12.
Solberg HE, Gräsbeck R. Reference values. Adv Clin Chem 1989;27:1-79.
Solberg HE, Lahti A Detection of outliers in reference distributions: performance of Horn's algorithm. Clin Chem 2005;51(12):2326-32.
Solberg HE, Stamm D. IFCC recommendation: The theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values. J Automat Chem 1991;13(5):231-234.
Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver-operating characteristic analysis. Clin Chem 2003;49(3):433-439.
Walton RN. Establishing reference intervals: Health as a relative concept. Seminars in Avian and Exotic Pet Medicine; 2001; 10(2):66-71.
Wiinberg B, Jensen AL, Kjelgaard-Hansen M, et al. Study on biological variation of haemostatic parameters in clinical healthy dogs. Vet J 2007;174:62-68.

Zweig, MH. Assessment of the clinical accuracy of laboratory tests using receiver operating characteristic (ROC) plots; Approved Guideline, Jan. 1995. CLSI/NCCLS Document GP10-a. NCCLS, 940 W Valley Road, Suite 1400, Wayne, Pennsylvania 19087, USA

Figure 1. Procedural steps for *de novo* determination of RI for new methods or new populations.

1. Perform literature search for information about analytes to be measured (preliminary investigation).
2. Define reference population and establish selection, inclusion and exclusion criteria (Section 1 and Table 1).
3. Develop questionnaire to be completed by examining clinician, owner/caretaker or both in order to determine if reference individual fits the selection or partitioning criteria (Section 2).
4. Determine number of reference individuals available or the number required to establish reference intervals with desired level of certainty (as reflected by 90% CI around the reference limits) (Section 3).
5. Select reference individuals, preferably by direct methods (Section 4).
6. Collect and handle reference samples in standardized manner (Section 5).
7. Analyze reference samples using well-controlled methods (Section 6 and 7).
8. Prepare histogram (Section 8).
9. Identify outliers (Section 9). This may require prior transformation to appropriately apply outlier detection methods and may need to be repeated after initial outliers are eliminated.
10. Determine distribution of reference data (Gaussian or non-Gaussian) (Section 10). If using parametric methods, transform data if it is not Gaussian and retest distribution. Transformation may improve the performance of the robust method. Nonparametric methods do not require any particular distribution.
11. Calculate upper and lower reference limits using an appropriate statistical method based on distribution of data and number of samples (Section 11 and Table 2). Calculate confidence intervals for the upper and lower reference limits.
12. Determine the need for partitioning only if there are sufficient numbers of reference samples or there is evidence for clinical importance (Section 12).
13. Document all previous steps for a comprehensive reference interval summary report (Section 13 and Table 3).

List of definitions

These terms appear roughly in the order in which they appear in the text of the guidelines document. This was done in order to group similar terms together.

1. **Reference intervals (RI)**. An interval contains all the possible values between, and including, an upper and lower limit. **Reference limits** are defined such that the reference interval contains a specified proportion of values from a reference population. Because the term 'range' refers to a single number representing the number of values between 2 limits, reference ***interval*** is preferred over reference ***range***.
2. A **reference population** is an undefined number of individuals that represent the demographic for which the reference intervals will be used. **Reference individuals** are chosen, preferably at random, from this larger population to provide **reference samples** for the establishment of a reference interval. The numerical results derived from these samples are referred to as **reference values**.
3. **De novo reference intervals** "De novo" means "from the beginning," "afresh," "anew," or "beginning again. This term refers to RI established by a specific laboratory from reference samples that were collected expressly for this purpose.
4. **Selection criteria** define the desired characteristics of a reference individual. The specific criteria chosen will depend of the purpose of the reference interval and the specific population the RI is intended to represent.
5. **Exclusion criteria** are defined so that individuals that should not be included in the reference sample population are excluded.
6. **Partitioning criteria** are used to further subdivide a reference population into a more refined demographic. **Partitioning** creates narrower RI and may be used when there are important biological differences that impact measureable quantities in the partitioned subgroups.

   Example: Selection criteria: healthy adult cats; Exclusion criteria: cats< 6 months of age, signs of illness; Partitioning criteria: gender
7. **Direct and indirect sampling methods. Direct sampling** methods involve selecting healthy individuals from a general population and collecting blood samples from them in order to generate results. **Indirect sampling** methods involve selecting results from a medical database and utilizing statistical methods to eliminate values that appear to come from unhealthy individuals.
8. ***A priori* and *a posteriori* sampling methods.** These terms refer to timing of the application of selection criteria. In ***a priori*** sampling methods, individuals are selected according to predefined criteria followed by collection of samples. This method is used when there is sufficient information about the biological quantity. In ***a posteriori*** sampling, samples are collected from individuals and only after the results are known are selection criteria applied. This latter method typically is used when little prior information is known about the biological quantity.

9. **Analytical error** is composed of **random** (CV) and **systematic** (bias) **error**. **Random error** (also call **imprecision**) refers to the variation between repeated measurements on the same sample. **Systematic error** (also called **inaccuracy**) refers to the difference between the measurement of a quantity and its true value. The true value may be defined by analysis using a gold-standard method.
10. **Coefficient of variation (CV)** describes the error around the mean presented as a proportion of the mean; CV = SD/mean.
11. A **histogram** provides a graphical representation of the distribution of reference data. The value (or concentration) of the measurable quantity is plotted in intervals along the x-axis and the frequency of measurements within that interval on the y-axis. It is the preferred method for visually presenting reference data and can be used to initially estimate the distribution of the data as well as to tentatively identify outliers.
12. **Gaussian** describes reference data that is normally distributed around the mean such that 95% of the reference values fall within 2 standard deviations of the mean.
13. **Outliers** are values that do not belong to the underlying distribution of the data. Outliers may result from erroneous inclusion of results from an individual that did not satisfy the selection criteria (e.g., inclusion of results from a diseased individual). Outliers also may results from other types of preanalytical, analytical and postanalytical error. True outliers can affect the location of the reference limits and should be identified and eliminated prior to calculating RI.
14. **Type I and Type II error.** Type I error is the rejection of the Null hypothesis when it is true. Type II error is the acceptance of the Null hypothesis when it is false. With regards to the question of whether a certain reference value is an outlier, Type I error indicates the elimination of a proposed outlier when it should be included and Type II error indicates the acceptance a proposed outlier when it should be eliminated.
15. **Transference** refers to the adoption by a laboratory of previously established RI. Procedures for **validation of RI** must be completed by the adopting laboratory prior to the use of the transferred RI to ensure that they are appropriate to the laboratory's patient population and laboratory methods.
16. **Normal deviate test and the z-statistic.** The **normal deviate test** is a statistical test used to determine whether the means of 2 populations are significantly different. The **z-statistic** is a standardized scoring tool that indicates how many standard deviations an observation is above or below the mean. These statistical tools are used to determine the need for partitioning and require the data to have a Gaussian distribution.
17. **Binomial test** is statistical test used to query data within 2 categories. It asks whether the proportion of data that falls within each category occurred by chance

or for some predetermined reason. A simplified version of the binomial test can be used to validate a transferred RI.

18. **Biological variability** refers to the variation in a measureable quantity between individuals.
19. **Individual or subject-based reference intervals** are reference intervals derived from a single individual. These may be useful when a sufficient number of reference individuals cannot be collected to create valid population-based RI, and when high biological variability limits the usefulness of population-based RI to detect important changes in an individual patient.
20. **Reference change value** (also called critical difference) is the difference between consecutive measurements of an analyte in an individual that is considered significant ($p \leq 0.05$). This is calculated based on known biologic variation within a species and analytical imprecision of the instrument used for analysis of samples.
21. A **decision limit** is a pre-determined threshold which distinguishes between 2 populations, e.g., those with a specific disease and those without the disease. Decision limits are defined by consensus and based on investigations of animals with and without a specific disease.

Table 1. Criteria for the selection, partitioning or exclusion of reference individuals

Selection criteria (may be used as partitioning criteria)		
Biological	Age	Example: neonate, juvenile, adult
	Sex	Example: female, male, altered
	Breed	Example: Holstein, Angus
Clinical	History	Example: no signs of illness in the 2 weeks preceding or following sample collection
	Preventative health care	Example: vaccination, routine anthelmintics
	Health	Example: Physical exam
	Diagnostic evaluation	Example: routine hematology, biochemistry, urinalysis; imaging studies
	Husbandry	Example: farmed, free-living, diet
Geographical		Examples: coastal, temperate, mountain, specific state or region, ambient temperature
Exclusion criteria (*may serve as partitioning criteria)		
Biological	Metabolic	Example: fasted or non-fasted, intense exercise, high stress
	Cell damage	Example: traumatic venipuncture, physical or chemical restraint
Physiologic		Examples: illness, medications, lactation*, pregnancy*
Medications		Examples: hormones or growth promoters, enzyme inducers (corticosteroids, antileptics)

Table 2. Recommended procedures for establishing RI based on reference sample size and distribution

Sample size	Data distribution (innate or transformed)	Statistical method
≥ 120	Not applicable	Nonparametric with 90% CI of ref. limits
40 ≤ x <120	Gaussian	Robust with 90% CI of ref. limits Parametric with 90% CI of ref. limits
	Non-Gaussian	Robust with 90% CI (preferred) of ref. limits Nonparametric[a]
20 ≤ x <40	Gaussian	Parametric with 90% CI of ref. limits[b]
	Non-Gaussian	Robust with 90% CI of ref. limits[b]
10 ≤ x < 20	Not applicable	Do not calculate reference intervals[b]
< 10	Not applicable	Do not report reference values

Confidence interval (CI)

[a]Cannot determine 90% CI nonparametrically with <120 reference sample, alternative methods required, e.g., bootstrap.

[b]Include the following information: histogram, mean or median, minimum and maximum

Table 3. Information to include in the RI study document or when publishing RI studies.

Item	Explanation
Demographics of reference population	Geographic location Source of reference individuals/samples Species and breed(s) Number of individuals from which samples were collected Age and gender distribution Husbandry (housing, diet, vaccines, parasite control, etc.) Determinants of health status Other details if pertinent
Preanalytical methods	Patient preparation Sample collection method (tube type, etc.) Sample handling and processing Time/season of collection if pertinent
Analytical methods	Analyzer (make and model) Methodology and reagents Quality specifications (TEa, bias, CV) Quality control reagents and procedures
Method of data analysis	Histogram Outlier identification method Reasons for eliminating certain values Evaluation of distribution Definition of interval (e.g. central 95%, 2.5^{th} and 97.5^{th} percentile limits) Number of reference samples (n) used to determine RI Method of interval determination (e.g., parametric, nonparametric, robust) 90% confidence intervals of the reference limits
Additional information	Raw data from reference samples Date RI implemented in the laboratory Date RI retired from use Dates of re-evaluation or re-validation of RI