Veterinary Clinical Pathology ISSN 0275-6382

SPECIAL REPORT

ASVCP reference interval guidelines: determination of de novo reference intervals in veterinary species and other related topics

Kristen R. Friedrichs[1], Kendal E. Harr[2], Kathy P. Freeman[3], Balazs Szladovits[4], Raquel M. Walton[5], Kirstin F. Barnhart[6], Julia Blanco-Chavez[7]

[1]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin–Madison, Madison, WI, USA; [2]URIKA, LLC, Mukilteo, WA, USA; [3]IDEXX Laboratories, Wetherby, West Yorkshire, UK; [4]Department of Pathology and Infectious Diseases, The Royal Veterinary College, Hertfordshire, UK; [5]Laboratory of Pathology and Toxicology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA; [6]Keeling Center of Comparative Medicine and Research, MD Anderson Cancer Center, Bastrop, TX, USA; and [7]All Creatures Veterinary Clinic and Lodge, Rittman, OH, USA

Key Words
Common RI, decision limits, laboratory standards, quality assurance, subject-based RI, transference

Correspondence:
Dr. Kristen R. Friedrichs, Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin–Madison, 2015 Linden Dr., Madison, WI 53706, USA
E-mail: friedrik@vetmed.wisc.edu

DOI:10.1111/vcp.12006

Abstract: Reference intervals (RI) are an integral component of laboratory diagnostic testing and clinical decision-making and represent estimated distributions of reference values (RV) from healthy populations of comparable individuals. Because decisions to pursue diagnoses or initiate treatment are often based on values falling outside RI, the collection and analysis of RV should be approached with diligence. This report is a condensation of the ASVCP 2011 consensus guidelines for determination of de novo RI in veterinary species, which mirror the 2008 Clinical Laboratory and Standards Institute (CLSI) recommendations, but with language and examples specific to veterinary species. Newer topics include robust methods for calculating RI from small sample sizes and procedures for outlier detection adapted to data quality. Because collecting sufficient reference samples is challenging, this document also provides recommendations for determining multicenter RI and for transference and validation of RI from other sources (eg, manufacturers). Advice for use and interpretation of subject-based RI is included, as these RI are an alternative to population-based RI when sample size or inter-individual variation is high. Finally, generation of decision limits, which distinguish between populations according to a predefined query (eg, diseased or non-diseased), is described. Adoption of these guidelines by the entire veterinary community will improve communication and dissemination of expected clinical laboratory values in a variety of animal species and will provide a template for publications on RI. This and other reports from the Quality Assurance and Laboratory Standards (QALS) committee are intended to promote quality laboratory practices in laboratories serving both clinical and research veterinarians.

Position Statements and Special Reports developed by the American Society for Veterinary Clinical Pathology (ASVCP) provide current information on topics in veterinary clinical pathology that are important to the veterinary community. The procedure for submitting statements is detailed at www.asvcp.org/membersonly/positionpapers.cfm. The ASVCP Executive Board is responsible for the review and approval of all statements, often following a period of input from the ASVCP membership and experts in the field. The final draft is then submitted to *Veterinary Clinical Pathology* and is edited prior to publication.

Introduction

The concept of population-based reference values (RV) was introduced in human medicine in 1969 and subsequently applied to veterinary species.[1-6] Reference values are typically reported as reference intervals (RI) comprising 95% of a healthy reference population. Since their introduction, population-based RI have become one of the most common tools used in clinical decision-making processes. Although use of popula-

tion-based RI is universally accepted, the optimal method for their derivation is frequently debated and is a recurring topic in the clinical laboratory literature.

The standard for the production of human population-based RI was commissioned by the International Federation of Clinical Chemistry (IFCC) in 1970 and resulted in a 6-part series that was adopted by several professional organizations, including the Clinical and Laboratory Standards Institute (CLSI).[7-12] A revision released by the CLSI in 2008 includes recommendations for transference and validation of RI from other sources and for the use of robust methods for determining RI from small sample sizes.[13]

The American Society for Veterinary Clinical Pathology (ASVCP) has recommended adherence to CLSI guidelines for determination of population-based RI. However, guidelines specifically addressing veterinary species would have numerous benefits to the veterinary medical community. In response to this need, the Quality Assurance and Laboratory Standards (QALS) committee of the ASVCP formed a subcommittee to generate guidelines for de novo determination of RI in veterinary species. In addition, the subcommittee addressed the related topics of transference and validation, multicenter and subject-based RI, and establishment of decision limits. Because RI are specific to a particular set of conditions, the guidelines also briefly address the misuse of published RI. This report represents an abridged edition of the guidelines found in their entirety on the ASVCP website.[14]

These guidelines, modeled after the revised CLSI document[13] summarized in a recent review,[15] were evaluated and adopted by the ASVCP membership in November 2011. Intended users include professionals working in veterinary diagnostic laboratories where assays for clinical and research purposes are performed on animal samples, manufacturers of veterinary diagnostic equipment and assays, and authors of articles on RI in veterinary species. In addition, veterinary clinicians who use RI should be familiar with these procedures so they can evaluate RI studies to ensure appropriate application to their patient population. Familiarity with relevant definitions and terms used in RI studies is also advised (Table 1).

Determination of de novo Reference Intervals for New Analytes, New Methods, or New Populations

Preliminary investigation

Investigation into sources of biological variability and interferences affecting measurement of the analytes (or variables or measurands) 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 is used to establish inclusion and exclusion criteria and determine the need for separate RI based on animal factors or preanalytical techniques. 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-fasted condition, or poor sample handling.

Selection of reference individuals

The reference population and criteria used to establish health must be defined. Demographics of the reference population should represent the animal population for which the RI will be used. Selection or inclusion criteria that describe individuals to be included and verify health should be established. Selection criteria include biological, clinical, and geographical factors (Table 2).[16] For laboratory animal species, origin and strain may also have significant effects on the accuracy of RI. In a recent study, it was concluded that RI for laboratory animals obtained from different suppliers can be highly variable, necessitating reverification of RI if animals are obtained from a new supplier.[17] Procedures to verify health may be narrowly defined (history and physical examination) or extensive (complete minimum database, imaging, fecal examination for parasites, and clinical follow-up) depending on the source of reference individuals and specifications of the RI study. Exclusion criteria should be defined to identify animals that should not be included (Table 2).[18] Establishing health in wild-caught species is especially challenging, and protocols should be rigorously established to minimize variation caused by including subjects of indeterminate health. Some inclusion and exclusion criteria may also function as partitioning criteria (eg, age, sex, or reproductive status), which permit more refined RI within subgroups.[16]

For RI established from clinical populations, a questionnaire should be created that establishes whether a potential reference individual conforms to selection criteria, belongs to a partitioned subgroup, or should be excluded. The questionnaire is completed by the owner or caretaker and by the individual(s) examining the subject and collecting the specimens. Owner consent for participation in the RI study is often included and is mandatory at many institutions.

Consideration of the number of individuals available to provide reference samples should be addressed

Table 1. Glossary of terms relevant to reference interval (RI) studies.

Term	Definition
Analytical error	Analytical error is composed of random (coefficient of variation [CV]) and systematic (bias) error. Random error (also called imprecision) refers to the variation between repeated measurements of 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. Coefficient of variation describes the error around the mean presented as a proportion of the mean; CV = SD/mean.
Binomial test	The binomial test is a statistical test used to query data within 2 categories. It asks whether the proportion of data that fall within each category occurred by chance or for a predetermined reason. A simplified version of the binomial test can be used to validate a transferred RI.
Biological variability	Biological variability refers to the variation in a measureable quantity between individuals.
de novo RI	de novo means 'from the beginning', 'afresh', or 'beginning again'. This term refers to RI established by a specific laboratory from reference samples that were collected expressly for this purpose.
Decision limit	A decision limit (decision threshold) is a predetermined threshold that distinguishes between 2 populations, eg, those with and without a specific disease. Decision limits are defined by consensus and based on investigations of animals with and without a specific disease.
Exclusion criteria	Exclusion criteria are defined to eliminate individuals that should not be included in the reference sample population.
Gaussian	Gaussian describes reference data that are normally distributed around the mean such that 95% of the reference values fall within 2 SDs of the mean.
Histogram	A histogram provides a graphical representation of the distribution of reference data. The quantity (concentration or activity) of the measurand (analyte) is plotted in intervals on 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 data as well as to tentatively identify outliers.
Normal deviate test	The normal deviate test is a statistical test used to determine whether the means of 2 populations are significantly different.
Outliers	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 selection criteria (eg, inclusion of results from a diseased individual). Outliers also may result from other types of preanalytical, analytical, and postanalytical error. Outliers affect reference limits and should be identified and eliminated prior to calculating RI.
Partitioning criteria	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 have an impact on measureable quantities in the partitioned subgroups.
Reference change value	Reference change value (also called critical difference) is the difference between consecutive measurements of an analyte in an individual that is considered significant ($P \leq .05$) and is calculated based on known biological variation within a species and analytical imprecision.
Reference interval	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. Reference interval is preferred over the term reference range.
Reference population	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 establishment of a reference interval. The numerical results derived from these samples are referred to as reference values.
Sampling methods (direct, indirect, a priori, a posteriori)	Direct sampling methods involve selecting healthy individuals from a general population and collecting samples in order to generate results. Indirect sampling methods involve selecting results from a medical database and utilizing statistical methods to eliminate values from presumed unhealthy individuals. In 'a priori' sampling, 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 selection criteria applied only after the results are known; this method typically is used when there is little prior information about the biological quantity.
Selection criteria	Selection criteria define the desired characteristics of a reference individual. The specific criteria will depend on the purpose of the reference interval and the specific population the RI will represent.
Subject-based RI	Subject-based (or individual) RI are derived from a single individual and may be useful when a sufficient number of reference individuals cannot be collected to create valid population-based RI or when high biological variability limits the usefulness of population-based RI to detect important changes in an individual patient.
Transference	Transference refers to adoption of previously established RI by a laboratory. Procedures for validation of RI must be completed by the adopting laboratory prior to use of the transferred RI to ensure that the RI are appropriate to the laboratory's animal patient population and laboratory methods and quality.
Type I and Type II error	Type I error (α) is rejection of the null hypothesis when it is true. Type II error (β) is acceptance of the null hypothesis when it is false. Type I error indicates the elimination of a proposed outlier when it should be included, and Type II error indicates the acceptance of a proposed outlier when it should be eliminated.

(continued)

Table 1. (continued)

Term	Definition
z-statistic	The *z*-statistic is a standardized scoring tool that indicates how many SDs above or below the mean an observation is. The normal deviate test and *z*-statistic are statistical tools used to determine the need for partitioning and require the data to have a Gaussian distribution.

Table 2. Criteria for the selection and exclusion of reference individuals.

Classification	Category	Example
*Selection criteria**		
Biological	Age	Neonate, juvenile, adult
	Sex	Female, male, altered
	Breed	Holstein, Angus
	Strain	Cynomolgus monkey from Mauritius or Southeast Asia
	Supplier	Sprague-Dawley rats from 2 different suppliers
Clinical	History	No signs of illness in the 2 weeks preceding or following sample collection
	Preventative care	Vaccination, routine anthelmintics
	Health	Physical examination
	Diagnostic evaluation	Routine hematology, biochemistry, urinalysis, imaging studies
	Husbandry	Farmed, free-living, diet
Geographic	Location	Coastal, temperate, mountain, specific state or region
	Environment	Ambient temperature
Exclusion criteria		
Biological	Metabolic	Fasted or non-fasted, intense exercise, high stress
	Cell damage	Traumatic venipuncture, physical or chemical restraint
Physiologic		Illness, medications, lactation*, pregnancy*
Medications		Hormones or growth promoters, enzyme inducers (corticosteroids or antileptics)

*May be used as partitioning criteria.

early in the RI study. A minimum of 120 reference individuals is recommended in order to determine reference limits by nonparametric methods with 90% confidence intervals (CI). Additional samples should be collected to allow for possible rejection of outliers. Reference intervals determined from smaller sample sizes are commonplace and often unavoidable in veterinary medicine. However, thorough consideration of the effect of small sample size on the accuracy and precision of population-based RI should be addressed early in the study. The smaller the sample size is, the higher is the degree of uncertainty in the estimation of reference limits.

Direct sampling methods are preferred over indirect sampling methods. Indirect sampling may include results from unhealthy individuals and consequently may not truly reflect the distribution of analyte measurements in a healthy population. When using direct sampling methods, inclusion and exclusion criteria may be applied a priori or a posteriori depending of the amount of information available for the analyte in question (see Table 1 for definitions).

Preanalytical procedures: patient preparation, sample collection, and analytical quality

Preparation of reference individuals, sample collection, sample handling, and sample processing should be performed in a standardized manner that is consistent with the methods used for testing of animal patients.[9] 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 (Table 3). Sample type (eg, serum or plasma) should be the same for all reference samples. Details regarding preanalytical factors should be included in the RI study document.

Estimates of analytical quality (coefficient of variation [CV] and bias or allowable total error [TE_A]) should be recorded for all methods. These may be determined during the RI study or during prior

Table 3. Preanalytical factors for consideration and standardization based on prior knowledge of the analyte and effects these factors have on sample quality and results.

Preanalytical factor	Examples
Patient preparation and handling	Fasted or non-fasted Method of capture and restraint Use of sedation or general anesthesia
Sample collection	Site (jugular, cephalic, or tail vein; sublingual artery) Preparation of site Anticoagulant Collection system (syringe and needle, vacutainer)
Sample handling	Transportation Temperature (room temperature, on ice) Time to clot Centrifugation Anaerobic
Time of collection	Circadian rhythms and seasonal fluctuations, especially for hormones
Analyte stability	Storage conditions and duration; determine if samples can be analyzed in batches or must be analyzed immediately

method validation (MV). Estimates should fall within the acceptable quality requirement goals of the laboratory. Quality goals may be based on biological variation, clinical interpretation of test results, consensus documents, or a combination of these.[19,20] Information on quality performance is necessary for RI to be considered for transference.

Analytical procedures

Samples should be analyzed using methods that are stringently monitored with appropriate quality control procedures.[21] Conditions for analysis should be well defined in a manner consistent with analysis of animal patient samples in order to reduce variation that is not due to inter- or intra-individual variability. However, variation that is part of everyday operations, such as changes in reagent lots and technical staff, should be integrated into RI studies whenever possible to approximate normal working conditions. The laboratory should establish a submission policy for RI study samples and criteria for rejection due to poor sample quality. Results should be monitored in real-time so that errors can be detected when re-measurement is possible; this will prevent excessive rejection of RV by reducing the number of potential outliers. Details of analytical methods, including the make and model of the analyzer as well as the source of reagents and quality control materials, should be recorded.

Statistical analysis of reference values

Graphical presentation reference data

Statistical analysis of RV begins with the preparation and examination of histograms that illustrate distribution of the data as well as highlight potential outliers. Histograms are preferred over box-plots or dot-plots for displaying reference data.

Identification and elimination of outliers

Identification and elimination or correction of outliers are important steps in the evaluation of reference data. Outliers are RV that do not belong to the underlying distribution and include extreme values resulting from inadvertent inclusion of samples from unhealthy or non-representative individuals or those affected by preanalytical (eg, poor sample quality), analytical, or postanalytical (eg, transcription) errors. Reference values affected by these types of errors should be eliminated whether they lie in the extremities of the distribution or not. The presence of extreme outliers has a detrimental effect on determination of reference limits, especially when calculated parametrically.[22] When reference individuals are selected randomly from well-defined populations and health is confidently established, retention of all RV is favored. However, when reference individuals are selected by convenience, health is not readily confirmed (eg, wild-caught species), or field methods introduce higher levels of inaccuracy and imprecision, RV located at the extremities should be examined more rigorously for possible exclusion.

Identification of outliers begins by examining the histogram; however, values lying at the extremities of the distribution should not be eliminated indiscriminately. Appropriate statistical methods should be used to verify these values as outliers. The 2 most commonly used outlier detection methods in RI studies are Dixon's range statistic[23] and Horn's algorithm using Tukey's interquartile fences[24]; however, other methods are available.[25,26] Selection of an appropriate outlier detection method depends on several factors. Some methods require that reference data have Gaussian distributions, which may require prior transformation (Tukey's interquartile fences). When this prerequisite is not followed, values located in the tails of skewed data may be erroneously eliminated. The presence of multiple outliers located at one or both extremities may have a masking effect and render certain methods unsuitable for outlier detection. As noted above, when reference data are likely to be contaminated with potential outliers (eg, data from

wild-caught species), methods more likely to exclude values should be selected (eg, selection of a Dixon's range statistic from a table of critical values with confidence levels of $\alpha = .1$ instead of $\alpha = .05$).[27] If outlier detection methods cannot be applied appropriately, nonparametric methods, which are less influenced by the presence of outliers, are preferred for determining RI. Clinical experience also should be used when determining when to retain or eliminate certain values. Because these statistical methods may not identify all outliers,[28] the best ways to avoid inclusion of inappropriate values (outliers) within the reference data are to ensure that all reference individuals are healthy and belong to the desired demographic and to avoid unintended preanalytical and analytical variation by adhering to study protocols.

Dixon's outlier range statistic typically examines the single most extreme value and favors retention (a conservative approach).[23] The simplest criterion of rejection (called the r criterion or r statistic) is $D/R > 0.3$, where D is the absolute difference between the most extreme value and the next nearest value, and R is the range of all values including the extreme value.[29] If more than one outlying value is observed at one extremity, the least most extreme value can be treated as the most extreme for calculation of the ratio (called a block procedure). If this value is identified as an outlier, then all the more extreme values can be eliminated.[30] To more readily identify and eliminate values as outliers, the r criterion can be compared to tables of critical values based on the number of anticipated outliers (one- or 2-tailed), the number of RV, and the desired level of confidence (a liberal approach).[27] Horn's algorithm using Tukey's interquartile fences identifies multiple outliers located at the upper and lower extremities and generally favors elimination (a liberal approach).[24] The criterion for rejection is values exceeding interquartile (IQ) fences set at $Q_1 - 1.5*IQR$ and $Q_3 + 1.5*IQR$, where IQR is the interquartile range, $IQR = IQ_3 - IQ_1$, and IQ_1 and IQ_3 are the 25th and 75th percentiles, respectively. Once outliers have been eliminated, data should be queried again for additional outliers, which may have been unmasked. Finally, the number of outliers eliminated and the reason for exclusion should be recorded in the RI study document.

Determination of distribution

Distribution of reference data may be Gaussian or non-Gaussian. Because statistical methods for identifying outliers and determining RI may be distribution-dependent, the distribution of RV should be assessed by examining the histogram and confirmed with a goodness-of-fit test (Anderson–Darling, Kolmogorov–Smirnov, or Shapiro–Wilk). If the distribution is not Gaussian but parametric methods will be used, data should be transformed using an appropriate function (eg, log or Box-Cox transformation) and retested. Reference interval software programs typically include various transformation functions. If normality cannot be established, parametric methods cannot be used to establish RI. Nonparametric and robust methods do not require assumption of normality; however, robust methods perform better when data are symmetrically distributed.[13,31]

Determination of reference limits

By convention, RI encompass the central 95% of RV and are bounded by upper and lower reference limits. The statistical method selected for determination of reference limits is based on the number and distribution of RV (Table 4). Nonparametric methods are recommended when at least 120 RV are available. The 2.5th and 97.5th fractiles serve as the lower and upper reference limits, respectively; precise limits may be interpolated between 2 consecutive values.[24] Ninety-percent CI around these limits can be determined nonparametrically when at least 120 values are available. When fewer than 120 samples are available, alterna-

Table 4. Recommended procedures for establishing RI based on reference sample size and distribution.

Sample size	Data distribution (raw or transformed)	Statistical method
≥ 120	Not applicable	Nonparametric with 90% CI of reference limits
$40 \leq x < 120$	Gaussian	Robust with 90% CI of reference limits
		Parametric with 90% CI of reference limits
	Non-Gaussian	Robust with 90% CI of reference limits (preferred)
		Nonparametric*
$20 \leq x < 40$	Gaussian	Parametric with 90% CI of reference limits†
	Non-Gaussian	Robust with 90% CI of reference limits†
$10 \leq x < 20$	Not applicable	Do not calculate reference intervals†
< 10	Not applicable	Do not report reference values

*Cannot determine 90% CI nonparametrically with < 120 reference samples; alternative methods, eg, bootstrap, required.

†Include the following: histogram, mean or median, and minimum and maximum; alternatively, a table of all reference values can be provided along with the histogram.

CI indicates confidence interval.

tive methods (eg, bootstrapping) are required to determine 90% CI.[13] Thirty-nine is the minimum number of samples for which 95% nonparametric RI can be determined; however, in this case, the most extreme values serve as the lower and upper limits. If nonparametric methods must be used with small sample sizes, enough samples should be collected to allow trimming of potential outliers at both extremities. Alternatively, robust or parametric methods should be used.

Robust methods are recommended when $\geq$ 40 and $\leq$ 120 reference samples are available. The robust method utilizes an iterative process to estimate location and spread of data.[32,33] Although robust methods do not require Gaussian distributions, they perform better when data are distributed symmetrically. Bootstrap methods should be used to determine 90% CI. The robust method is included in several clinical laboratory software programs, including CBstat,[34] Reference Value Advisor freeware,[35,36] and MedCalc.[37]

Parametric methods may be used when $\geq$ 40 and $\leq$ 120 reference samples are available if the data have or can be transformed to Gaussian distributions. Parametric methods encompass slightly more than the central 95% of values and establish the upper and lower reference limits at mean + 2SD and mean − 2SD, respectively. Parametric methods should be used to determine 90% CI of the reference limits.

There will be instances in veterinary medicine when a limited number of reference samples can be collected, such as for special and wild-caught species and neonates. When $\geq$ 20 and < 40 reference samples are available, RI should be calculated by methods that are robust (distribution-independent) or parametric (if normality can be established). To highlight the uncertainty inherent with small sample sizes, 90% CI should be calculated. In addition, the following should be reported to allow informed clinical decision-making: histogram, mean or median, and minimum and maximum values. Alternatively, a table listing all the values can be provided along with the histogram.

When $\geq$ 10 and < 20 reference samples are available, a table of ascending values along with a histogram and mean or median values should be reported, but RI should not be determined given the uncertainty of limits based on so few samples. Reference data from fewer than 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. Subject-based RI should be considered when so few reference individuals are available (see section on biological variation, individuality, and subject-based RI). In general, when fewer than 40 reference individuals are available, emphasis should be 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, because the presence of a single outlier has a significant effect on the estimated reference limits. If outliers are identified, every effort should be made to collect replacement samples.

Confidence intervals

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 limits and are generally narrower for large samples sizes. Boyd and Harris recommend that CI should not exceed 0.2 times the width of the RI (WCI/WRI < 0.2, where WCI is the width of the CI and WRI is the width of the RI).[13] When CI exceed this limit, an effort should be made to collect additional reference samples.

Partitioning

Partitioning into subclasses should be based on physiologic differences that are expected to result in important clinical differences in RI. Partitioning favors homogeneous subpopulations, decreasing variability between individuals and narrowing the RI. However, partitioning should be considered only if there are at least 40 individuals within each subclass or if there are clear clinical reasons. Statistical criteria should consider not only subgroup means,[38] but also subgroup SDs.[39,40] Partitioning criteria also may examine the proportion of each subgroup that falls outside the upper and lower limits of a combined RI.[41,42] The following nonstatistical criteria support partitioning of RI: when descriptors used to assign an individual to a partitioned subgroup are easily obtainable, when reference limits serve as critical clinical decision limits, and when there are documented clinical differences between subgroups. Details regarding the application of statistical criteria for partitioning can be found in the online version of these guidelines[14] or in relevant references.[38-42]

Documentation

All previous steps and procedural details should be documented so that RI are clearly defined (Tables 5 and 6). A complete and detailed RI study document should be available to users upon request. The laboratory should retain RI summary documents for a predetermined amount of time or indefinitely. These details also should be included in publications of RI for veterinary species to allow critical evaluation by potential

Table 5. Procedural steps for de novo determination of reference intervals for new analytes, new methods, or new populations.

Step	Procedure
1	Perform literature search for information about analytes to be measured (preliminary investigation).
2	Define reference population and establish selection, inclusion, and exclusion criteria (Table 2).
3	Develop questionnaire to be completed by examining clinician, owner/caretaker, or both in order to determine if reference individuals fit the selection or partitioning criteria or should be excluded.
4	Determine number of reference individuals available or the number required to establish reference intervals with desired level of certainty (as reflected by 90% confidence intervals around the reference limits).
5	Select reference individuals, preferably by direct methods.
6	Collect and handle reference samples in standardized manner (Table 3).
7	Analyze reference samples using well-controlled methods.
8	Prepare histogram.
9	Identify outliers; 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). If using parametric methods, transform data if 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 (Table 4). 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.
13	Document all previous steps for a comprehensive reference interval summary report (Table 6).

Table 6. Information to include in the reference interval (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) or strain(s)
	Supplier (for laboratory animals)
	Number of individuals from which samples were collected
	Age and sex distribution
	Husbandry (eg, housing, diet, vaccines, parasite control)
	Determinants of health status
	Other details if pertinent
Preanalytical methods	Animal patient preparation
	Sample collection method (eg, anticoagulant)
	Sample handling and processing
	Time/season of collection if pertinent
Analytical methods	Analyzer (make and model)
	Methodology and reagents
	Quality specifications (eg, TEa, bias, CV)
	Quality control reagents and procedures
Method of data analysis	Histogram
	Outlier identification method
	Number of results eliminated and reasons for elimination
	Distribution of reference values (Gaussian or not)
	Definition of interval (eg, central 95%, 2.5th and 97.5th percentile limits)
	Number of reference samples (n) used to determine RI
	Method of interval determination (eg, 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 revalidation of RI

TEa indicates allowable total error; CV, coefficient of variation.

users. Spreadsheets for reporting reference data are available in an addendum to the RI Guidelines posted on the ASVCP website.[14] Reference intervals determined de novo within a laboratory should be reviewed every 3 to 5 years and revalidated if needed (see section on transference and validation).

Postanalytical procedures: laboratory presentation of reference intervals

Reference intervals are typically printed on the animal patient report, but this should only be done when they are applicable to that patient. It is useful to indicate which patient values are increased or decreased. Information that is valuable for clinical decision-making (eg, demographics of the reference population, preanalytical factors, and CI) but cannot be contained within the patient report should be available to the clinician upon request. Reference intervals that deviate from customary percentiles and limits or are specific to a certain subclass should clearly be identified on the report.

Alternatives to Establishing de novo RI

Establishing de novo RI is challenging, time-consuming, and expensive. Alternatives are described in this section and include transference and validation of RI adopted from other sources, common (or multicenter) RI, and subject-based RI.

Transference and validation of RI

In order to forego the expense and difficulty of establishing intra-laboratory RI, many laboratories adopt RI from other sources or from instrument manufacturers.[13] Reference intervals considered for transference should be derived from a similar animal patient demographic and collected under similar preanalytical conditions as the adopting laboratory. Transference is easier when methods are identical, but comparison of method procedures can be done if methods differ. If methods are comparable, RI can be transferred directly. If systematic differences (bias) exist, reference limits can be adjusted using regression statistics or difference of means.[13] Transference of RI between methods using regression statistics should be limited to a single occurrence. Instrument accuracy and precision and laboratory quality should be similar between the contributing and adopting laboratories. If significant differences exist in analytical quality, transference may not be appropriate.

Prior to utilization in a clinical setting, transferred RI, including those adopted from manufacturers, should be validated. The most direct validation method compares results from 20 healthy reference individuals collected from the laboratory's animal patient population with the proposed RI.[13] Results should be screened for outliers, and replacement samples collected if outliers are identified. Reference intervals for each analyte are considered valid if 0–2 results fall outside the proposed RI. If more than 4 results fall outside the proposed reference limits, the RI should be rejected and de novo RI determined. When 3 or 4 results fall outside the proposed RI, an additional 20 samples can be collected and interpreted as described previously. This method replicates the binomial test but will not determine whether the proposed reference limits are too wide for the laboratory's animal patient population (for example, if all 20 results fall within the proposed reference limits).[43]

A more rigorous validation utilizes advanced statistical methods (Mann–Whitney U test, median test, Siegal–Tukey test) to compare the proposed RI with results from 40–60 healthy reference individuals from the laboratory's animal patient population.[13] Alternatively, if 40–60 high-quality samples are available, de novo RI simply can be established using robust methods. Validation solely based on subjective assessment of the quality and applicability of the proposed RI is seldom sufficient, especially given the paucity of details concerning the origin of many RI currently in use.

Reference intervals should be revalidated every 3–5 years. In addition, revalidation is recommended when excessive false-positive and false-negative results are noted by clinicians and whenever there are significant changes in animal patient populations, preanalytical techniques, or analytical quality.[41]

A cautionary statement on the use of published reference intervals

Using published RI to interpret clinical laboratory results is common in veterinary medicine and is driven by the frequent lack of appropriate in-house RI for diverse animal patient populations. Interpreting clinical data using inappropriate RI may lead to misclassification of results and misdiagnosis, improper treatment, or both. Reference intervals published in textbooks, journal articles, and web-based databases may or may not contain sufficient information to determine whether the RI is appropriate for the clinic's or laboratory's animal patient population and preanalytical and analytical procedures. In addition, the quality and rigor with which the RI were established typically cannot be assessed. Published RI should be used with caution and only when sufficient information is available to determine their applicability. If published RI are adopted for extended use, appropriate validation procedures as described above should be performed.

Common (or multicenter) reference intervals

A second alternative to determining de novo RI within each laboratory is for several laboratories serving a similar animal patient population to contribute to the generation of common RI.[44,45] There are several advantages to this approach, not the least of which is the distribution of cost. Common RI are useful in mobile patient populations that may change locations and laboratories over a lifetime, as long as the population moves within the region represented by the common RI. In multi-laboratory RI studies, sample sizes often are much larger, providing greater certainty for the reference limits (narrower 90% CI) and allowing partitioning into subgroups based on age, sex, or other characteristics. Despite these advantages, establishing common RI requires a rigorous approach with particular attention to calibration and laboratory quality.

A prerequisite for common RI is similarity of the animal patient populations served by the participating laboratories. After this has been confirmed, uniform selection and exclusion criteria and sample collection and handling techniques should be established.

Although preferred, laboratories contributing results to a common RI do not have to use the same analyzer or methods. However, all analyzers must be calibrated to produce comparable results.[46] Calibrators should be either traceable international standards or a single pooled specimen; the assigned calibrator mean should be determined from means contributed by all participating laboratories with each laboratory determining its own mean from 10 to 20 replicates. Uniform quality requirement goals for imprecision (CV%) and inaccuracy (bias) should be established for all laboratories and may be based on biological variation, clinical interpretation of results, or both.[19,20] Bias in particular must be controlled and minimized by each laboratory to prevent misclassification of results. If a laboratory's bias exceeds established goals, procedures should be initiated to return bias to within acceptable limits (eg, recalibration). In order to maintain stable performance, assays should be rigorously monitored using quality control procedures designed for high probability of error detection and low probability of false rejection. Use of the same quality control materials facilitates standardization and comparison of results.

Biological variation, individuality, and subject-based reference intervals

Population-based RI serve as a comparison for animal patient test results when an alternative frame of reference is not available. However, due to relatively high inter-individual variability, population-based RI sometimes lack necessary sensitivity to detect changes in the health of an individual. In addition, it may be difficult to collect sufficient numbers of reference samples to generate representative population-based RI for some species (eg, zoologic and conservation species). In these instances, subject-based RI provide a viable alternative.[47]

Subject-based RI are most useful when intra-individual biological variation, represented by the mean CV for separate values from an individual (CV_I), is less than inter-individual variation, or the CV of a group (CV_G).[48] The index of individuality (II) provides an objective criterion to determine the relative utility of subject-based versus population-based RI. Index of individuality is calculated as $II = (CV_I^2 + CV_A^2)^{1/2}/CV_G$, where CV_A is analytical variation (random error or imprecision). Because $CV_A < CV_I$ for many automated analyzers, the II is often simplified as CV_I/CV_G.[48] When II is < 0.6, subject-based RI are preferable to population-based RI, whereas when it is > 1.4, individual RI yield no more information than traditional RI.[49]

When using subject-based RI, reference change values (RCV) serve to determine whether a difference between consecutive measurements in an individual is significant. The RCV is based on CV_I values in health and the dispersion of these variations across a population and is calculated as $RCV = z \times [2(CV_I^2 + CV_A^2)]^{1/2}$, where z represents the z-statistic, an estimate of probability.[50] RCV is optimally used when $CV_A < 0.5 \times CV_I$. Because $CV_A << CV_I$ for many automated assays, RCV can be simplified as $z \times 2^{1/2}CV_I$ (or $z \times 1.41CV_I$). The z-statistic conventionally used for RCV is $z = 1.96$, which provides a 50% probability of detecting an increase with a 5% probability of Type I error.[50] When larger z-statistics, such as 3.34, are used, the probability of a significant change being detected increases to 90% while increasing the probability of Type II error.[50]

To use subject-based RI, it is necessary to know CV_I for each analyte, as well as the imprecision (CV_A) of the analytical method. Published CV_I are available for many analytes in dogs[51,52] and in some exotic species.[53] Biological variation can be measured with only a small number of healthy reference individuals when care is taken to minimize preanalytical variation.[49] To monitor patients with chronic diseases, CV_I may need to be established for individuals with stable chronic disease instead of using those established in health.[54]

Establishing Decision Thresholds (Decision Limits)

Unlike RI that are descriptive of a population, decision limits discriminate between populations, typically those with and without disease. Reference intervals are determined statistically; decision limits are established experimentally and by consensus.[13] The ability to discriminate between patients with and without disease defines the accuracy of a diagnostic test and is a major determinant of its clinical utility. The accuracy of a test, in turn, depends on the decision limit selected for characterizing a test result as positive (diseased) or negative (non-diseased). Designing prospective studies to establish decision limits requires thorough deliberation and meticulous planning.[55]

The first step in designing a prospective study to establish decision limits is to define the clinical question or management decision to be made as well as the target population. The role that the diagnostic test will play in the decision-making process should be identified for a specific disease or condition. The target population should be characterized with regard to factors that may have an impact on interpretation of test results (eg, disease prevalence and duration and factors

such as age and sex). Next, individuals anticipated to have both positive and negative results for the test under investigation should be prospectively selected from the relevant target population. Study subjects should represent the expected diversity within the target population for test-positive and test-negative classifications. Selection of study subjects should be done independent of results for the test or analyte being evaluated. Statisticians can be consulted to establish sample sizes with sufficient power to yield statistically relevant results. True disease status should be established by comprehensive examination and testing procedures that may include standardized staging, grading, or scoring schemes or common outcome assessments.[56]

The test under investigation should be performed and interpreted by investigators who are blinded to the true disease classification of the subjects. If sample stability permits, samples can be assayed in batches to minimize between-run analytical variance. When comparing performance among multiple diagnostic tests, the test should be performed on all subjects at the same time or at the same stage of disease, and all tests should performed on the same sample. To avoid bias in favor of test performance, subjects with unexpected test results should not be excluded. Sensitivity and specificity are then calculated at a variety of decision thresholds and plotted on a receiver-operating characteristic (ROC) curve (x-axis, 1 – specificity; y-axis, sensitivity).[57,58] Optimal decision limits are selected by location on the ROC curve (most upper left point) or by determining the decision limit with the highest proportion of correct interpretations (most true-positive and true-negative results). The area under the curve (AUC) is an estimate of test accuracy. Under most circumstances, the more accurate test has a larger AUC. Confidence intervals can be determined around the points on an ROC curve. To provide quality assurance, true disease status should be monitored throughout the study and confirmed with long-term clinical follow-up or histopathologic examination of tissues collected at biopsy or necropsy.

Conclusions

A uniform and consistent approach to establishing RI will benefit the entire veterinary medical community. Implementation of statistical methods for establishing RI from small sample sizes, the acceptance of RI utilizing transference and validation, 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. Compliance with these guidelines by professionals who establish and publish RI for animals should facilitate communication within the broader veterinary medical community. By providing detailed information in RI studies, 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. Finally, the use of decision limits will grow as the number of available tests and our understanding of specific disease states expands.

Acknowledgments

The authors would like to thank Drs. J.P. Braun (École Nationale Vétérinaire de Toulouse) and Mads Kjelgaard-Hansen (University of Copenhagen) for their expert review of this document and Drs. Chet Thomas (University of Wisconsin–Madison) and Joanne Paul-Murphy (University of California–Davis) for sharing their reference data from wild-caught lizards.

References

1. Gräsbeck R, Saris NE. Establishment and use of normal values. *Scand J Clin Lab Invest*. 1969;26:S62–S63.
2. Lumsden JH, Mullen K. On establishing reference values. *Can J Comp Med*. 1978;42:293–301.
3. Lumsden JH, Mullen K, McSherry BJ. Canine hematology and biochemistry reference values. *Can J Comp Med*. 1979;43:125–131.
4. Lumsden JH, Mullen K, Rowe R. Hematology and biochemistry reference values for female Holstein cattle. *Can J Comp Med*. 1980;44:24–31.
5. Lumsden JH, Rowe R, Mullen K. Hematology and biochemistry reference values for the light horse. *Can J Comp Med*. 1980;44:32–42.
6. Friendship RM, Lumsden JH, McMillan I, Wilson MR. Hematology and biochemistry reference values for Ontario Swine. *Can J Comp Med*. 1984;48:390–393.
7. 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.
8. PetitClerc C, Solberg HE. Approved recommendation (1987) on the theory of reference values. Part 2. Selection of individuals for the production of reference values. *Clin Chim Acta*. 1987;170:S1–S11.
9. Solberg HE. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation of individuals and collection of specimens for the pro-

duction of reference values. *Clin Chim Acta*. 1988;177:S1–S12.

10. 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;13:231–234.
11. 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.
12. 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.
13. Clinical and Laboratory Standards Institute. *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guidelines*. 3rd ed. Wayne, PA: CLSI; 2008.
14. ASVCP Quality Assurance and Laboratory Standards Committee (QALS) Guidelines for the Determination of Reference Intervals in Veterinary Species and other related topics. Available at: http://www.asvcp.org/pubs/pdf/RI%20Guidelines%20For%20ASVCP%20website.pdf. Accessed September 17, 2012.
15. Geffré A, Friedrichs K, Harr K, Concordet D, Trumel C, Braun JP. Reference values: a review. *Vet Clin Pathol*. 2009;38:288–298.
16. Walton RM. Establishing reference intervals: Health as a relative concept. *Sem Avian Exotic Pet Med*. 2001;10:66–71.
17. Kampfmann I, Bauer N, Johannes S, Moritz A. Differences in hematologic variables in rats of the same strain but different origin. *Vet Clin Pathol*. 2012;41:228–234.
18. Poole T. Happy animals make good science. *Lab Anim*. 1997;31:116–124.
19. Kenny D, Fraser CG, Hyltoft Petersen P, Kallner A. Consensus agreement. *Scand J Clin Invest*. 1999;59:585.
20. Kjelgaard-Hansen M, Jensen AL. Subjectivity in defining quality specifications for quality control and test validation. *Vet Clin Pathol*. 2010;39:133–135.
21. Flatland B, Freeman KP, Friedrichs KR, et al. ASVCP quality assurance guidelines: control of general analytical factors in veterinary laboratories. *Vet Clin Pathol*. 2010;39:264–277.
22. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and nonhealthy individuals on reference interval estimation. *Clin Chem*. 2001;47:2137–2145.
23. Dixon WJ. Processing data for outliers. *Biometrics*. 1983;9:74–89.
24. Horn PS, Pesce AJ. Reference intervals: an update. *Clin Chim Acta*. 2003;334:5–23.
25. Grubb FE. Sample criteria for testing outlying observations. *Annal Math Stat*. 1950;21:27–58.
26. Jain RB. A recursive version of Grubb's test for detecting multiple outliers in environmental and chemical data. *Clin Biochem*. 2010;43:1030–1033.
27. 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;63:139–146.
28. Solberg HE, Lahti A. Detection of outliers in reference distributions: performance of Horn's algorithm. *Clin Chem*. 2005;51:2326–2332.
29. Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate or normal range. *Clin Chem*. 1971;17:275–284.
30. Barnett V, Lewis T. *Outliers in Statistical Data*. Chichester, UK: John Wiley and Sons, Ltd.; 1978:381–386.
31. Horn PS, Pesce AJ. *Reference Intervals: A User's Guide*. Washington, DC: AACC Press; 2005:47–57.
32. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. *Clin Chem*. 1998;44:622–631.
33. Horn PS, Pesce AJ, Copeland BE. Reference interval computation using robust vs. parametric and nonparametric analyses. *Clin Chem*. 1999;45:2284–2285.
34. CBstat. Available at: http://direct.aacc.org/ProductCatalog/Product.aspx?ID=2475. Accessed September 22, 2011.
35. Reference Value Advisor. Available at: http://www.biostat.envt.fr/spip/spip.php?article63. Accessed September 22, 2011.
36. Geffré 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.
37. MedCalc. Available at: http://www.medcalc.org/index.php. Accessed September 22, 2011.
38. 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.
39. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranged. *Clin Chem*. 1990;36:265–270.
40. Lahti A, Hyltoft Petersen P, Boyd JC, Fraser C, Jørgensen N. Objective criteria for partitioning Gaussian distributed reference values into subgroups. *Clin Chem*. 2002;48:338–352.
41. Ceriotti F, Hinzmann R, Panteghini M. Reference intervals: the way forward. *Ann Clin Biochem*. 2009;46(Pt 1):8–17.
42. 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;50:891–900.

43. Horn PS, Pesce AJ. *Reference Intervals: A User's Guide*. Washington DC: AACC Press; 2005:77–80.
44. Jones GRD, Barker A, Tate J, Lim C, Robertson K. The case for common reference intervals. *Clin Biochem Rev*. 2004;25:99–104.
45. Hyltoft Petersen P, Rustad P. Prerequisites for establishing common reference intervals. *Scand J Clin Lab Invest*. 2004;64:285–292.
46. Jensen AL, Kjelgaard-Hansen M. Method comparison in the clinical laboratory. *Vet Clin Pathol*. 2006;35:276–286.
47. Walton RM. Subject-based reference values: biological variation, individuality, and reference change values. *Vet Clin Pathol*. 2012;41:175–181.
48. Fraser CG. Inherent biological variation and reference values. *Clin Chem Lab Med*. 2004;42:758–764.
49. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. *Crit Rev Clin Lab Sci*. 1989;27:409–437.
50. 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.
51. Jensen AL, Aaes H. Critical differences of clinical chemistry parameters in blood from dogs. *Res Vet Sci*. 1993;54:10–14.
52. 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.
53. 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;38:217–221.
54. Ricós C, Iglesias Canadell N, García-Lario J, et al. Within-subject biological variation in disease: collated data and clinical consequences. *Ann Clin Biochem*. 2007;44:343–352.
55. Clinical and Laboratory Standards Institute. *Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline*. Wayne, PA: CLSI; 1995.
56. 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.
57. 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.
58. Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem*. 2003;49:433–439.