To cite this version :
GREFFE, Anne, BRAUN, Jean-Pierre, TRUMEL, Cathy, CONCORDET, Didier. Estimation of reference intervals from small samples: an example using canine plasma creatinine. *Veterinary Clinical Pathology,* 2009, vol.38, n°4, p. 477-484. ISSN 0275-6382.

Estimation of reference intervals from small samples: an example using canine plasma creatinine

A. Geffré[1], J.P. Braun[1,2], C. Trumel[1], D. Concordet[2,3]

[1]Department of Clinical Sciences, [2]UMR181 Physiopathologie and Toxicologie Expérimentales INRA, ENVT, and [3]Department of Biological Sciences, Ecole Nationale Vétérinaire, Toulouse, France

Key Words
Creatinine, dog, reference values, robust method, small sample

Background: According to international recommendations, reference intervals should be determined from at least 120 reference individuals, which often are impossible to achieve in veterinary clinical pathology, especially for wild animals. When only a small number of reference subjects is available, the possible bias cannot be known and the normality of the distribution cannot be evaluated. A comparison of reference intervals estimated by different methods could be helpful.
Objective: The purpose of this study was to compare reference limits determined from a large set of canine plasma creatinine reference values, and large subsets of this data, with estimates obtained from small samples selected randomly.
Methods: Twenty sets each of 120 and 27 samples were randomly selected from a set of 1439 plasma creatinine results obtained from healthy dogs in another study. Reference intervals for the whole sample and for the large samples were determined by a nonparametric method. The estimated reference limits for the small samples were minimum and maximum, mean ± 2 SD of native and Box–Cox-transformed values, 2.5th and 97.5th percentiles by a robust method on native and Box–Cox-transformed values, and estimates from diagrams of cumulative distribution functions.
Results: The whole sample had a heavily skewed distribution, which approached Gaussian after Box–Cox transformation. The reference limits estimated from small samples were highly variable. The closest estimates to the 1439-result reference interval for 27-result subsamples were obtained by both parametric and robust methods after Box–Cox transformation but were grossly erroneous in some cases.
Conclusion: For small samples, it is recommended that all values be reported graphically in a dot plot or histogram and that estimates of the reference limits be compared using different methods.

Introduction

The concept of using reference values for reporting the variability of analytes in healthy subjects is widely accepted as a basis for interpreting the individual values observed in patients, even though many medical classifications are based on decision limits or consensus values that differ from the reference limits.[1,2] The most recent international guidelines for the preparation of reference limits in human clinical pathology have been published by the International Federation of Clinical Chemistry (IFCC) and the Clinical and Laboratory Standards Institute (CLSI).[3] Most of these recommendations can be transposed to animal clinical pathology. However, as with some human subgroups (eg, newborns or elderly people), it is often impossible to obtain the minimum recommended number (120) of reference subjects. It is nevertheless recommended that in such cases "data should still be analyzed by the nonparametric method. As an alternative, the robust method may be used . . .."[3]

Robust methods are based on iterative processes that estimate the median and spread of the distribution.[4,5] In the IFCC-CLSI proposed guideline for robust methods, they gave as an example the assessment of a reference interval for plasma calcium in women, in which 3 sets of 20 values each were randomly selected from a group of 120 values. The subsequent reference intervals calculated by the robust method were very close to the interval determined by a nonparametric method from the whole set of 120 values.[3] In that example, the distribution of 120 values was roughly Gaussian and the range was narrow (88–103 mg/L), with no outliers. This may explain why "the performance of the method (was) not dependent on getting a 'good' set of points; though of course, results vary depending on the specific values selected." Despite the good results obtained in that example, in the final approved revised guideline the working group was hesitant to recommend calculating reference intervals with sample numbers < 80 "except in the most extreme instances."[3]

In practice, when only a small number of reference subjects are available for selection, the possible bias resulting from this selection cannot be known, and the normality of the distribution cannot be evaluated. Thus some doubts will remain and a comparison of the reference intervals estimated by different methods could be helpful.

The aim of this study was to take a large set of canine plasma creatinine reference values obtained in a previous study[6] and to: (1) randomly select large samples (120-sample sets, which is the smallest number of subjects recommended for use of the nonparametric method) and determine reference intervals; (2) randomly select small samples (27-sample sets) and estimate the reference limits in each by different methods; and (3) compare the results obtained with the reference interval determined for the whole sample by the nonparametric method.

Materials and Methods

The whole sample consisted of 4097 dogs, of which 1439 were healthy animals of known body-weight class and plasma creatinine concentration. These healthy dogs consisted of 800 small (< 15 kg), 261 medium (15–35 kg), and 378 large (> 35 kg) dogs. No other possible factors of variation were taken into account in the study. Ten sets of 120 results each (large samples) were randomly selected from the whole sample using the ALEA function in Excel (Microsoft Corporation, Redmond, WA, USA). Ten additional large subsets consisting of 67 small, 22 medium, and 31 large dogs were selected to reflect the relative proportion of each body-weight class in the whole sample. Reference limits and their 90% confidence intervals (CIs) were determined for the whole population, for the random 120-sample subsets, and for the 120-sample subsets balanced for body weight, by the nonparametric method according to IFCC recommended procedure, which is based on the ranking of values and setting limits at the 0.025 and 0.975 fractiles (percentiles).[7]

Ten sets of 27 results each (small samples) were randomly obtained from the whole sample (R subgroup) by the same random selection method and 10 additional sets of 27 results were obtained by randomly selecting results from 15 small, 5 medium, and 7 large dogs, again to represent their relative proportion in the whole sample (S subgroup). In each of the 20 sets of 27 values, the following were calculated: (1) minimum–maximum interval (range); (2) mean ± 2 SD of native and Box–Cox-transformed values; (3) 2.5–97.5% intervals by a robust method with native and Box–Cox-transformed values; and (4) 2.5–97.5% intervals estimated graphically from the cumulative distribution functions (CDF) derived from the histograms.

Statistical analysis

Calculations were performed with an Excel spreadsheet (Microsoft Corporation) and the Analyse-It (Leeds, UK) set of macroinstructions. The Box–Cox λ coefficient was calculated with freeware R 2.7.0 (The R Foundation for Statistical Computing; http://stat.ethz.ch/CRAN/). The robust method used was the one recommended by Horn and colleagues.[3,4] Results were tested using ANOVA. Comparisons between groups was made using a Mann–Whitney *U*-test after testing the homogeneity of variances, and when necessary using Bonferroni's correction. Possible partitioning criteria were studied by Harris and Boyd's *z*-statistics,[8] and outliers were detected by visual inspection of distributions and confirmed by Tukey's criterion.[3]

Results

Reference limits of the whole reference sample ($n = 1439$)

The overall distribution was skewed and non-Gaussian (Figure 1) and could not be transformed into a Gaussian distribution by log or Box–Cox transformation (Anderson–Darling, $P = .0005$ and .001, respectively). The reference interval for plasma creatinine concentration determined nonparametrically (90% CI of limits in parentheses) was 53.1 (52.0–55.0) to 150.4 (148.0–159.0) μmol/L (Table 1). The effect of body weight was highly significant (ANOVA, $P < .001$) and

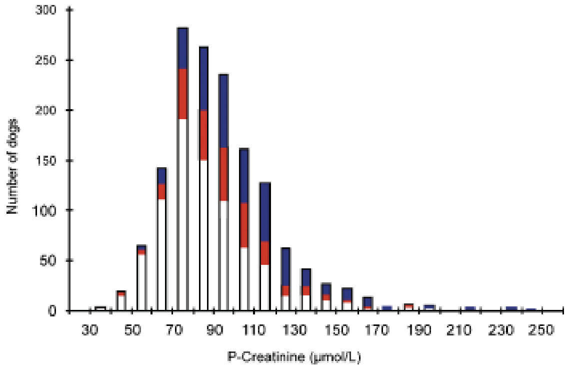

Figure 1. Distribution of plasma (P) creatinine concentration in 1439 clinically healthy dogs. White, body weight (BW) < 15 kg; red, BW 15–35 kg; blue, BW > 35 kg.

the 2 × 2 differences between the 3 body-weight classes were also significant (Mann–Whitney with Bonferroni correction, $P < .001$; Harris and Boyd $z \gg z^*$ after Box–Cox-transformation). The corresponding reference intervals for small, medium, and large dogs, respectively, were 51.0 (46.0–53.0) to 146.0 (140.0–153.0) µmol/L, 60.1 (53.1–62.0) to 143.9 (136.0–154.0) µmol/L, and 67.5 (61.9–70.0) to 168.6 (159.0–180.0) µmol/L.

Reference limits of the large samples

The distributions of all 120-sample subsets differed significantly from Gaussian (Anderson–Darling, $P < .05$) whereas none of the Box–Cox-transformed distributions, except 1 ($P = .014$), differed from Gaussian (Anderson–Darling, $P = .053–.977$). There was no difference between the limits determined by the nonparametric method in the randomly selected and body-weight class selected subgroups (Student's *t*-test after testing for homogeneity of variances, $P = .250$ and .829 for lower and upper limits, respectively). Lower and upper limits ranged from 44 to 57 (median 53.1) µmol/L and from 140 to 179 (median 150.7) µmol/L, respectively, and the 90% CIs ranged from 35 to 63 µmol/L and from 124 to 242 µmol/L for the lower and upper limits, respectively (Table 1 and Figure 2).

Reference limits of the small samples

The percentages of dogs in the 3 body-weight classes in the 10 randomly selected data subsets (R1–R10 subgroups) differed considerably from those in the whole sample (55.6%, 18.1%, 26.3%), whereas the percentages in the 10 S-subgroups, selected based on body-weight class, were almost the same (55.6%, 18.5%, 25.9%) as in the whole sample (Figure 3). Creatinine concentrations in the 20 subsets had similar ranges, with coefficients of variations (CVs) of 18.3–37.5% in the R subgroups and 18.2–39.4% in the S subgroups (Figure 4).

In most cases, the native values could not be used to estimate the reference limits by the robust method so the results are not reported. Extrapolation of the 2.5th and 97.5th percentiles from the histograms was imprecise due to the distribution of values (see 2 typical examples in Figure 5) and systematically gave values below the observed minimum and maximum (Wilcoxon's test, $P < .001$). Whatever the method used to calculate reference limits, no difference was found between the limits determined for the R and S subgroups, except for the mean ± 2 SD limit, which was

Table 1. Experimental approach and resulting reference limits for plasma creatinine concentration (µmol/L) in dogs.

				Reference Interval Results for Creatinine (µmol/L)	
Dataset	Method of Estimation	Data Type	Reference Interval Type	Lower Limit	Upper Limit
Whole sample ($n = 1439$)	Nonparametric (2.5–97.5%)	Native values	Limit (90% CI of limit)	53.1 (52–55)	150.4 (148–159)
Large subsets of 120 samples each	Nonparametric (2.5–97.5%)	Native values	Median (range) of limits	53.1 (44–57)	150.7 (140–179)
Small subsets of 27 samples each	Minimum–maximum	Native values	Median (range)	53.1 (40–75)	150.4 (125–239)
	Parametric (mean ± 2 SD)	Native values	Median (range)	42.0 (22–61)	146.1 (124–186)
		Box–Cox transformed	Median (range)	53.8 (40–74)	154.3 (126–283)*
	Robust	Native values	Median (range)	NC	NC
		Box–Cox transformed	Median (range)	52.4 (23–73)	165.8 (122–363)†
	Visual estimation	CDF	Median (range)	46.8 (30–70)	143.2 (120–223)

*Upper limit is an outlier; next nearest value is 203 µmol/L.

†Upper limit is an outlier; next nearest value is 215 µmol/L.

CDF, cumulative density distribution CI, confidence interval; NC, not calculated (not possible to calculate) for many subgroups.

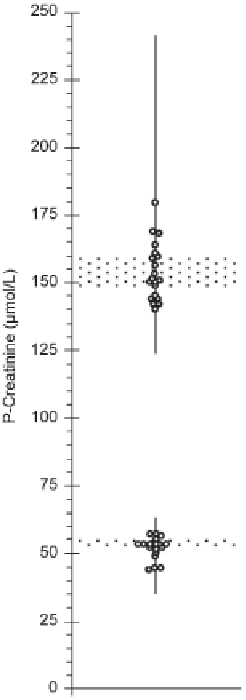

Figure 2. Upper and lower limits of the reference interval for plasma (P) creatinine determined by a nonparametric method in 20 randomly selected 120-sample subsets within the full dataset of 1439 healthy dogs. Dotted areas, 90% confidence intervals (CIs) of the limits determined for the whole sample; vertical bar, range of the 90% CI calculated from the 120-sample subsets.

lower in the S subgroups (Mann–Whitney, $P < .05$ after testing for homogeneity of variances).

Estimates of the lower limit of the reference interval obtained by the parametric and the robust method after Box–Cox transformation (Table 1 and Figure 6) were closer to the value determined in the whole dataset than those obtained by other methods. The range of upper limit estimates was wider than for the lower limit (Table 1 and Figure 7). Box–Cox-transformed results revealed an apparent outlier, which, according to Tukey's criterion, was eliminated. When this outlier was removed, the CV of the estimates was similar to that obtained for the lower limit. Closest estimates of the upper limit determined in the whole sample were obtained from the calculated mean ± 2 SD using native values and Box–Cox-transformed values.

The means of all estimates of the reference limits (50 and 158 µmol/L) were close to the values determined from the whole sample. The range of values was large, however, independent of the method used; the range was 39.8–73.6 µmol/L in the best case for the lower limit, and most values calculated from small

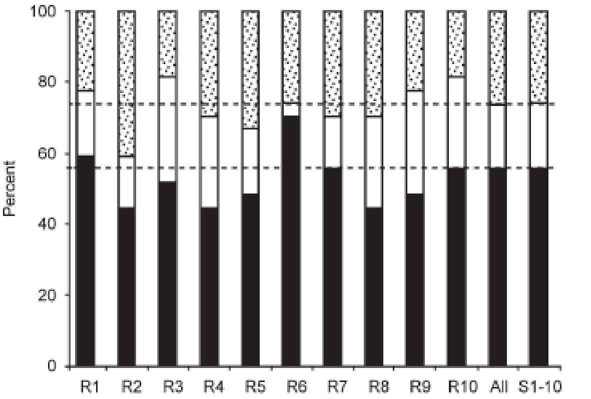

Figure 3. Percentage of small (black), medium (white), and large (speckled) dogs in the 10 sets of 27 results selected randomly (R1–R10) from the whole sample (All) and in the 10 sets of 27 results selected according to body-weight class (S1–10). Dotted lines are the percentages of dogs in each body-weight class in the whole sample.

samples were outside the CIs determined from the whole sample. Meaningful calculation of the CIs for the determined limits was precluded by the low number of individuals in the 20 small sets.

Discussion

The aim of this study was to determine whether reference limits calculated from small samples randomly selected from a large sample were identical to those calculated from the latter by nonparametric method. If so, more or less valid reference intervals could be estimated from small samples when large samples are not available. Our results, however, suggest the bias obtained from different random small samples resulted in

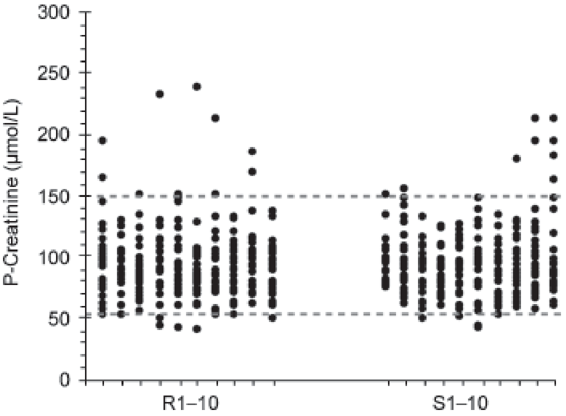

Figure 4. Distribution of plasma (P) creatinine values in the 10 sets of 27 results randomly selected (R1–R10) and in the 10 sets of 27 results selected according to body-weight class (S1 -10) from the whole sample. The dotted lines are the limits of the reference interval determined from the whole sample.

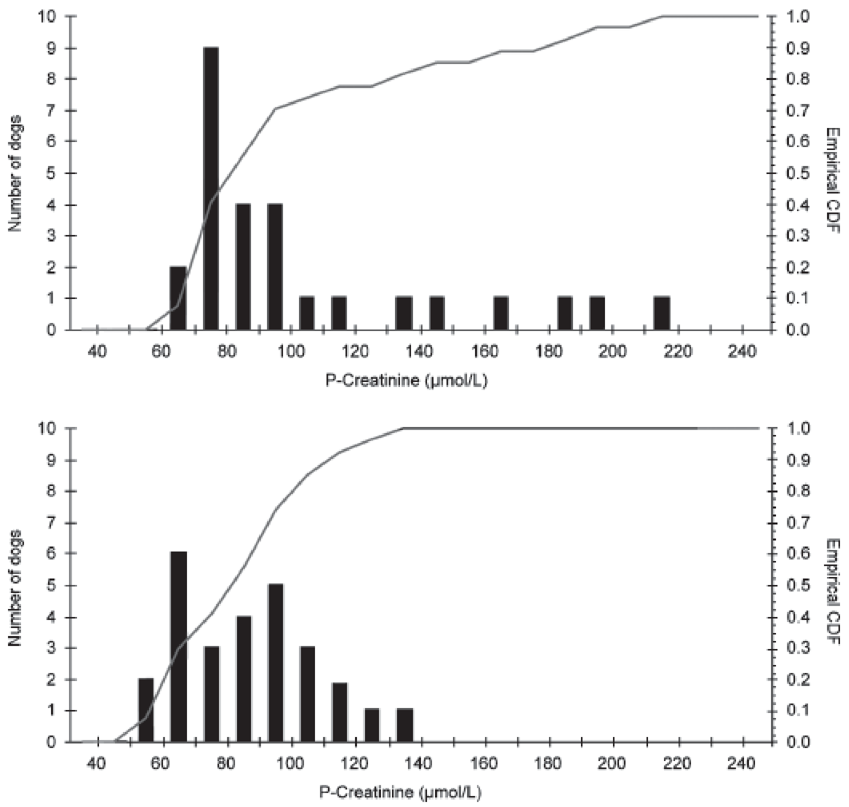

Figure 5. Examples of histograms (black bars) and cumulative distribution function (CDF; thin line) of plasma (P) creatinine values obtained from two 27-sample subsets of values.

very different estimates of reference limits compared with those obtained with large samples.

It was presumed that all recommended criteria concerning preanalytical and analytical criteria were respected.[3,9] These were not reported here as the aim was not to determine reference intervals for plasma creatinine concentration in dogs, but rather to compare different methods of estimating reference intervals from small samples. Even though some criteria may have been inadequate, their effect on the whole

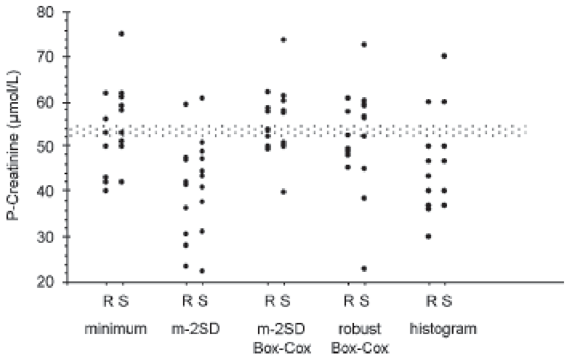

Figure 6. Estimates of the lower limits of the reference interval for plasma (P) creatinine as determined by different methods in the 10 sets of 27 results selected randomly (R) and in the 10 sets of 27 results selected according to body-weight class (S). The dotted area is the 90% confidence interval of the lower limit determined from the whole sample. m, mean; SD, standard deviation.

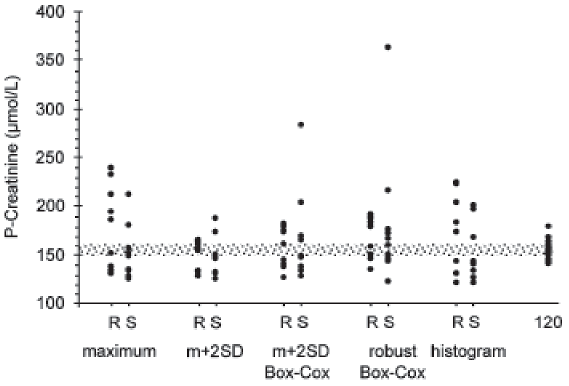

Figure 7. Estimates of the upper limit of the reference interval for plasma (P) creatinine as determined by different methods in the 10 sets of 27 results selected randomly (R) and in the 10 sets of 27 results selected according to body-weight class (S). The dotted area is the 90% confidence interval of the upper limit determined from the whole sample. m, mean; SD, standard deviation.

sample and the subgroups would have been identical and therefore validate the comparisons.

No apparent heterogeneity could be detected from the shape of the histogram of plasma creatinine concentration in the whole sample, except that it was heavily skewed toward high concentrations, which suggested that plasma creatinine concentrations might have been higher in 1 subgroup of dogs. Not all the possible partitioning factors evidenced in the preceding study were taken into account.[6] Body weight was the only partitioning factor used to investigate possible effects on results obtained from small samples, as this was reported previously to influence canine plasma creatinine concentration.[10,11] It is usually acknowledged that differences between subgroups may be statistically significant, when a large number of samples is compared, even though they may not be clinically relevant, which was the case in this study.[12] When applied to medical data, tests of partitioning such as that described by Harris and Boyd[8] are only suitable for comparisons of 2 sets of values. In the case of body weight, all 2 × 2 comparisons were highly significant. Thus, partitioning results according to the 3 body-weight classes was considered relevant, and validated the original selection of data subsets based on body-weight distribution.

The overall upper limit of the reference interval in this study was almost the same as the limit reported in the previous study in which all healthy animals were included (151 μmol/L; $n = 1516$ values).[6] In the present study, only those cases with known body-weight class were analyzed ($n = 1439$).

Although it was not the main aim, the variability of reference intervals determined using the recommended nonparametric method and the minimum number of reference individuals (120) was examined and compared with the variability of the estimates obtained from small numbers. The mean of the range of reference limits for the 20 random large samples selected from the whole sample and that of the reference limits determined from the whole sample, was almost identical. The variance of the limits thus determined was almost the same (they differed by ~7%) for the lower and upper limits. This was much lower than the variability of estimates from the small samples, but much broader than the 90% CIs determined for the whole sample. The narrowness of this latter was due to the large number of values used, uncommon in studies of reference values except those based on hospital data.[13] It was surprising, however, to see that in some cases the calculated limits derived from the minimum recommended number of 120 samples differed notably from those determined from the whole sample.

Twenty-seven was chosen as the number of samples for the small subsets in this study because many reports of reference values for nondomestic animal species include < 30 animals and because this number permitted a relative weighting of body-weight classes proportional to that of the whole sample.

In this study we confirmed that the bias obtained from different random samplings of 27 results resulted in very different estimates of reference limits but the results for the R and S subgroups, ie, with or without a partitioning factor (here, body-weight class) did not differ. Partitioning factors, based on a priori estimates of possible effects of sex, age, season, etc, on the results, are sometimes taken into account in studies with small numbers of animals. This approach may therefore not be appropriate if the number of animals in each category is too small to allow a proper study of differences between subgroups.

When only a small sample is available, recommendations state that all of the results may "serve a useful clinical purpose as a guide in the form of a list of all the values, [. . .] ordered according to increasing magnitude."[7] The major advantage of this type of data presentation is that no information is lost as all values are reported, but the list of numbers is not easy to evaluate. The values can also be reported in a dot plot, histogram, or diagram of CDF; these forms of data presentation may be more useful from a clinical standpoint. However, such lists or figures still are less easy to apply routinely than the upper and lower limits of a reference interval. When the number of samples is low, extrapolation of values from a CDF diagram resulted in 2.5% and 97.5% limits that were below the observed minimum and maximum of the dataset and therefore are not relevant. A nonparametric approach cannot be used as the number of reference samples is below 40.

Reference intervals from small samples are reported in the literature in many different ways, including median and maximum–minimum values, mean ± SD, and 2.5th and 97.5th percentiles estimated by parametric methods, with or without transformation of the data. However, it is impossible to correctly assess the type of distribution in small samples and, as shown in this study, the shapes of the distributions and the numbers of apparent outliers differed considerably between the 20 sets of values. Thus it is difficult to make relevant decisions on the best mathematical model to apply. However, as blood analyte distributions are often skewed, it would seem reasonable to include all the values obtained from small samples.

In this study, the lower reference limits were underestimated when the mean − 2 SD interval from untransformed values was used on the small samples, which would be expected from a distribution skewed toward high values. It was surprising, however, that the mean+2 SD limits were close to the upper limit calculated from the whole sample. The parametric approach is more valid if a better fit to Gaussian distribution can be obtained by mathematical transformation. This can often be achieved by Box–Cox transformation, which is now readily available in the R 2.7.0 freeware program. The distribution of the Box–Cox-transformed values did not differ significantly from Gaussian except for 1 subset of data, and parametric estimation of the reference interval from the Box–Cox-transformed values was close to the CIs of the reference limits determined from the whole sample. In the one exception, the upper limits determined as the mean+2 SD from Box–Cox-transformed values or by the robust method were grossly erroneous (Figure 7) compared with other methods of estimation.

Robust methods have been recommended for determining the quantiles of a distribution of small samples. In this case, the method of Horn et al[4] could not be applied to untransformed canine plasma creatinine values. As already mentioned, the robust method should preferably be used when the data fit a Gaussian distribution, and in this study, the robust approach was efficient on Box–Cox-transformed values. In the example in which the robust method was used for human plasma calcium values, estimation of reference limits was more accurate than in this study. This may be due to the higher interindividual variance of canine plasma creatinine (~15%)[14,15] than of human plasma calcium (~3%),[16,17] such that randomly selected small subsets of canine creatinine data may not be representative of the whole set of values, resulting in the determination of erroneous limits.

In summary, none of the methods used in this study were very satisfactory for estimating reference intervals in small samples, and probably no method can be used that is generally applicable. Whatever calculations and mathematical models are applied, the limiting factor remains the a priori assumption that the small sample of values available is representative of samples to be tested in the future for diagnosis.[18] Reference intervals always should be estimated from the largest possible number of animals available so that nonparametric methods can be used to determine the reference limits and their CIs, and allow evaluation of possible partitioning factors. When only small samples are available, the estimation of reference intervals is biased by the sample, which may be more or less representative of the whole population, and this bias cannot be determined. In this case, as much information as possible should be reported, including lists of ordered values, dot plots, and/or histograms. A good approach when estimating reference intervals is to transform the data to obtain the best possible fit with Gaussian distribution and to compare the estimated limits obtained by parametric and robust methods. The real and estimated reference limits may differ considerably, demonstrating the bias inherent in reporting a single estimate for small samples.

References

1. Grasbeck R, Saris NE. Establishment and use of normal values. *Scand J Clin Lab Invest*. 1969;26(Suppl 110):62–63.
2. Petitclerc C. Normality: the unreachable star? *Clin Chem Lab Med*. 2004;42:698–701.
3. CLSI. *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline*. 3rd ed. Wayne, PA: CLSI; 2008.
4. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. *Clin Chem*. 1998;44:622–631.
5. Horn PS, Pesce AJ, Copeland BE. Reference interval computation using robust vs parametric vs nonparametric analyses. *Clin Chem*. 1999;45: 2284–2285.
6. Concordet D, Vergez F, Trumel C, et al. A multicentric retrospective study of serum/plasma urea and creatinine concentrations in dogs using univariate and multivariate decision rules to evaluate diagnostic efficiency. *Vet Clin Pathol*. 2008;37:96–103.
7. Solberg HE. Approved recommendation (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *J Clin Chem Clin Biochem*. 1987;25: 645–656.
8. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. *Clin Chem*. 1990;36:265–270.
9. Solberg HE, Stamm D. International Federation of Clinical Chemistry (IFCC) IFCC recommendation. The theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values. *Ann Biol Clin (Paris)*. 1991;49: 487–490.
10. Medaille C, Trumel C, Concordet D, et al. Comparison of plasma/serum urea and creatinine concentrations in the dog: a 5-year retrospective study in a commercial veterinary clinical pathology laboratory. *J Vet Med A Physiol Pathol Clin Med*. 2004;51:119–123.

11. Feeman WE III, Couto CG, Gray TL. Serum creatinine concentrations in retired racing Greyhounds. *Vet Clin Pathol*. 2003;32:40–42.

12. Lahti A. Partitioning biochemical reference data into subgroups: comparison of existing methods. *Clin Chem Lab Med*. 2004;42:725–733.

13. Solberg HE. Using a hospitalized population to establish reference intervals: pros and cons. *Clin Chem*. 1994;40:2205–2206.

14. Pagitz M, Frommlet F, Schwendenwein I. Evaluation of biological variance of cystatin C in comparison with other endogenous markers of glomerular filtration rate in healthy dogs. *J Vet Intern Med*. 2007;21:936–942.

15. Jensen AL, Aaes H. Critical differences of clinical chemical parameters in blood from dogs. *Res Vet Sci*. 1993;54:10–14.

16. Harris EK, DeMets DL. Biological and analytic components of variation in long-term studies of serum constituents in normal subjects. V. Estimated biological variations in ionized calcium. *Clin Chem*. 1971;17:983–987.

17. Lacher DA, Hughes JP, Carroll MD. Estimate of biological variation of laboratory analytes based on the third national health and nutrition examination survey. *Clin Chem*. 2005;51:450–452.

18. Horn PS, Pesce AJ. Reference intervals: an update. *Clin Chim Acta*. 2003;334:5–23.