Clin Chem Lab Med 2010;48(11):1537–1551 © 2010 by Walter de Gruyter • Berlin • New York. DOI 10.1515/CCLM.2010.319

IFCC Document

An appraisal of statistical procedures used in derivation of reference intervals

Kiyoshi Ichihara[1,*] and James C. Boyd[2] on behalf of the IFCC Committee on Reference Intervals and Decision Limits (C-RIDL)

[1] Yamaguchi University Graduate School of Medicine, Ube, Yamaguchi, Japan
[2] Department of Pathology, University of Virginia Health System, Charlottesville, Virginia, USA

Abstract

When conducting studies to derive reference intervals (RIs), various statistical procedures are commonly applied at each step, from the planning stages to final computation of RIs. Determination of the necessary sample size is an important consideration, and evaluation of at least 400 individuals in each subgroup has been recommended to establish reliable common RIs in multicenter studies. Multiple regression analysis allows identification of the most important factors contributing to variation in test results, while accounting for possible confounding relationships among these factors. Of the various approaches proposed for judging the necessity of partitioning reference values, nested analysis of variance (ANOVA) is the likely method of choice owing to its ability to handle multiple groups and being able to adjust for multiple factors. Box-Cox power transformation often has been used to transform data to a Gaussian distribution for parametric computation of RIs. However, this transformation occasionally fails. Therefore, the non-parametric method based on determination of the 2.5 and 97.5 percentiles following sorting of the data, has been recommended for general use. The performance of the Box-Cox transformation can be improved by introducing an additional parameter representing the origin of transformation. In simulations, the confidence intervals (CIs) of reference limits (RLs) calculated by the parametric method were narrower than those calculated by the non-parametric approach. However, the margin of difference was rather small owing to additional variability in parametrically-determined RLs introduced by estimation of parameters for the Box-Cox transformation. The parametric calculation method may have an advantage over the non-parametric method in allowing identification and exclusion of extreme values during RI computation.
Clin Chem Lab Med 2010;48:1537–51.

*Corresponding author: Kiyoshi Ichihara, MD, PhD, Department of Clinical Laboratory Sciences, Faculty of Health Sciences, Yamaguchi University Graduate School of Medicine, Minami-Kogushi 1-1-1, Ube, 755-8505 Japan
Phone: +81-836-22-2884, Fax: +81-836-35-5213,
E-mail: ichihara@yamaguchi-u.ac.jp

Keywords: modified Box-Cox formula; multiple regression analysis; nested ANOVA; partitioning criteria; percentile; power transformation.

Introduction

Careful consideration of statistical methods and computational techniques is essential at each step of any study being performed to derive reference intervals (RIs). While planning the initial recruitment of healthy individuals as study participants, the necessary sample size must be determined appropriately, taking into account the desired reproducibility of the results. Following sampling of an appropriate size population of individuals and generating test results for each sample, the potential sources of variation need to be identified in order to interpret the test results and the RI properly. Statistical analysis is also necessary in studying these sources of variation to decide, whether separate reference values need to be estimated. Of the many statistical techniques proposed for the computation of RIs from the distributions of reference values, the optimal technique to be used depends on the characteristics of the underlying distribution of reference values. Once RIs have been derived using appropriate methods, the transference of those RIs to other laboratories may require additional statistical analysis.

In this article, we review various statistical procedures that can be employed in each step of a study for derivation of RIs and provide illustrations of each method. Comparisons among the various methods are included to demonstrate their advantages and disadvantages. The article is written with a perspective of the challenges encountered in analysis of large datasets, e.g., derived from the Asian multicenter project on reference values (manuscript in preparation).

This presentation has been structured to provide an introduction to these concepts at a fairly basic level. Several resources are available that provide more advanced treatments and details regarding many of the concepts that are presented below (1–3).

Sample size and standard error of reference limits

We begin our discussion by considering some elements in the design phase of a RI study. One of the first considerations is to determine how many reference individuals to study, i.e.,

determination of the necessary sample size. Usually, the necessary sample size is a function of how precisely the reference limits (RLs) need to be estimated. When the underlying distribution of test results in a healthy population follows a normal (or Gaussian) distribution, it is easy to estimate the 95% confidence intervals (95% CIs) of the RLs for both the lower limit (LL) and upper limit (UL), either by simulation or by analysis of differentials of the underlying normal distribution function (see below). The actual estimation of 95% CIs by a simulation study is shown in Figure 1.

In Figure 1, 95% CIs are shown across various sample sizes assuming that the underlying distribution has a mean of zero and a standard deviation (SD) of 1.0. To generate this plot, a simulation study was performed in which a sample of fixed size between 40 and 2000 was randomly generated 2000 times. The RI for each sample was computed parametrically as the mean $\pm$ 1.96SD, giving a 95% CI with endpoints, LL and UL, delimiting the central 95% of the population, [1] without transformation and [2] after power transformation (see below). In the latter case, the LL and UL are reverse transformed to obtain the RI endpoints in the original scale. The RI is also computed non-parametrically as [3] the 2.5th–97.5th percentile range. For these three methods, the corresponding central value was computed as [1] the mean, [2] the reverse transformed mean, or [3] the 50th percentile, respectively. The 95% CIs of the mean and UL by each of the three methods were estimated from the repeated trial for each data size. One-half the width of the 95% CI or one fourth of the RI width can be expressed as a parameter, s, which corresponds roughly to the between-individual SD (SD_{BI}), assuming an underlying Gaussian distribution with the mean of 0.0 and an SD of 1.0.

It is apparent that method [1] gives a much smaller 95% CI both for the mean and UL than methods [2] and [3], irrespective of the sample size. However, it must be noted that, in the real situation, we usually do not know the underlying data distribution. Thus, to compute RI by the parametric method, we need to include a step to normalize the distribution, usually using power transformation of the data. The large difference in 95% CI sizes between methods [1] and [2] is a reflection of the fact that the latter involves error in estimating the distribution parameters (described below), resulting in a larger uncertainty of the UL. However, the 95% CI by use of method [2] is narrower than that by method [3], but this margin is very small. Therefore, it is appropriate that the non-parametric method is recommended for derivation of the RI in the Clinical and Laboratory Standards Institute (CLSI) guideline (3). A possible advantage of the parametric method is its ability to identify and exclude extreme datapoints based on mean and SD of the distribution during computation of the RIs.

However, we should note in Figure 1 that the sample size of 120, described as the minimum size in the guideline, is far from optimal, since one-half of the 95% CI of the UL is

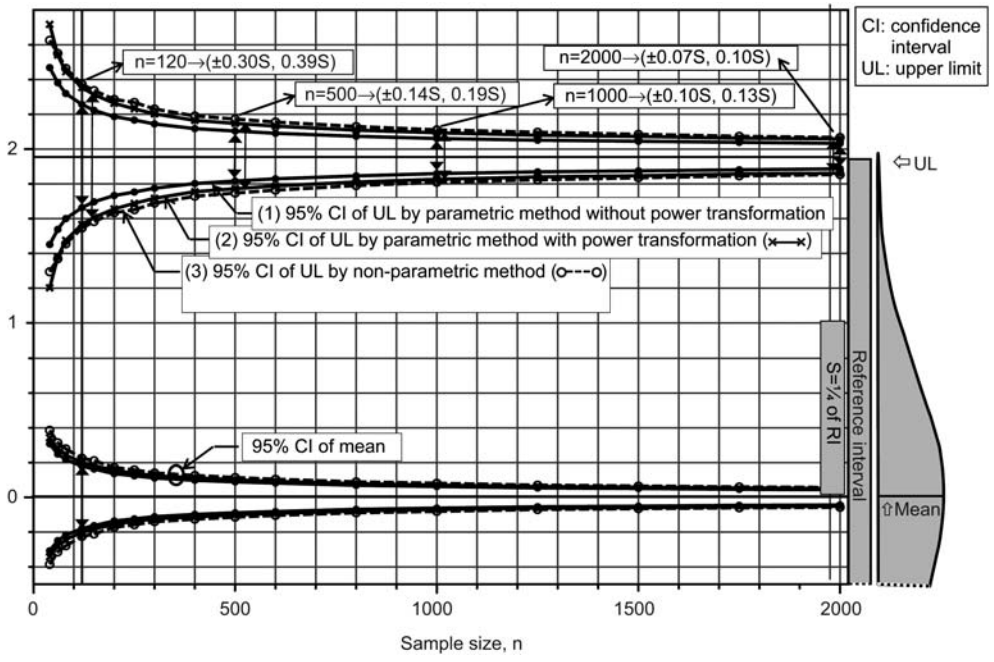

Figure 1 The 95% confidence intervals of reference limits in relation to the sample size and computation methods assuming a normal distribution.

By specifying a sample size between 40 and 2000, a reference sample consisting of normal random values with a mean of 0.0 and SD of 1.0 are generated 2000 times. The mean and UL of the distribution were computed (1) parametrically without power transformation, (2) parametrically with power transformation by use of modified Box-Cox formula, and (3) non-parametrically. The 95% confidence intervals of the mean and UL at each sample size are connected by lines to show their serial changes. The theoretical reference interval (truncated in its lower part) and magnitude of 1/4 of the width of RI, which corresponds approximately to the between individual SD is shown along the vertical axis of the figure.

$0.39 \times SD_{BI}$. If the RI is to be derived in a collaborative effort with participation of many laboratories, at least 400 subjects would be desirable (one-half of the 95% CI of the UL is $0.2 \times SD_{BI}$, which is nearly half of the above case and 5% ($=0.2/4=0.05$) of the width of the RI). If we set separate RIs for men and women, the minimal number of individuals needed would then be approximately 800. If we are seeking to derive RIs for each decade of life for individuals in their 20s, 30s, 40s, and 50s, separately for men and women, the needed sample size would be $800 \times 4 = 3200$ individuals.

As an alternative to the simulation method used above, the 95% CI of LL or UL for the parametric method without power transformation can be computed, assuming an underlying theoretical Gaussian distribution, using the following formula (1).

$$\textbf{95\% CI of LL(UL)} = \pm 2.81 \cdot \frac{s}{\sqrt{n}}. \quad (1)$$

When we assume, $s = 1.0$ as in the case shown in Figure 1, the 95% CI of UL ($\pm$one-half of the width) for $n = 120$ is ± 0.256; similarly, for $n = 500$, ± 0.126; for $n = 1000$, ± 0.089; for $n = 2000$, ± 0.062. These values roughly correspond to those of the simulation study shown in Figure 1 for method [1] in computing the RI.

These theoretical considerations all assume an underlying Gaussian distribution. In Figure 2, the 95% CIs for the LL, median, and UL of the RI are shown when a logarithmic normal distribution is assumed in the simulation. The distribution was rescaled so that the LL and UL pointed to –1.96 and 1.96, respectively, and 1/4 of the RI became approximately equal to 1.0. In Figure 2, an almost identical relationship as that in Figure 1 is shown between the CIs obtained by the parametric method with power transformation, and by the non-parametric method.

Analysis of sources of variations

Documentation of possible physiological sources of variations in test results among healthy individuals is necessary so that these sources can be taken into account in the interpretation of test results before considering results outside the RI as a reflection of possible underlying disease. Such information is especially important in judging the need for partitioning of reference values by any given factor. To assist in such an analysis during a RI study, a well designed questionnaire survey should be conducted at the time of specimen collection.

Univariate approaches

There are several statistical methods used to analyze sources of variation. Simple tests of differences among subgroups, partitioned by any given factor, can be performed by use of the t-test (or the Mann-Whitney U-test) between two sub-

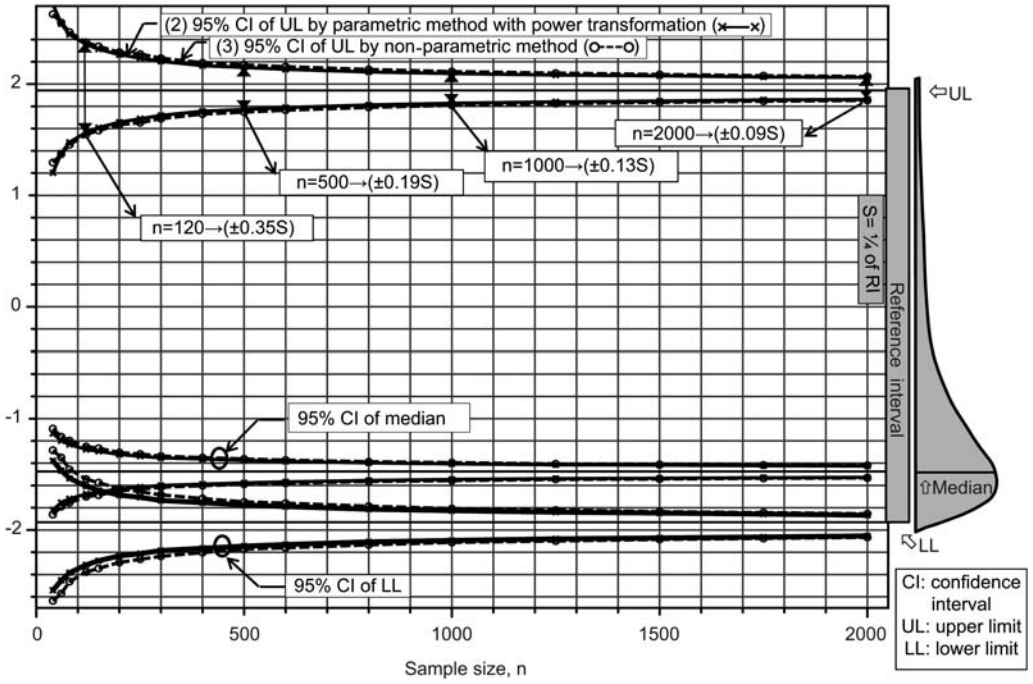

Figure 2 The 95% confidence intervals of reference limits in relation to the sample size and computation methods assuming a log-normal distribution.

The same scheme for simulation was used as in Figure 1 under the assumption of a log-normal distribution. The values are adjusted so that the theoretical RI for the data points to the interval between –1.96 and 1.96 and 1/4 of the RI became approximately equal to 1.0. The 95% confidence intervals for the LL, median, and UL are shown for (2) the parametric method with power transformation and for (3) the non-parametric method.

groups, or by use of a one-way ANOVA (or the Kruskal-Wallis test) when there are three or more subgroups. However, judgments based on univariate testing can be misleading when there are multiple factors influencing the test results.

Figure 3 illustrates such a problem we encountered in a multicenter study for derivation of RIs for serum proteins (4). In Figure 3A, reference values for serum immunoglobulin G (IgG) were partitioned according to smoking status. IgG concentrations were obviously higher in non-smokers. In Figure 3B, the same reference individuals were partitioned according to gender and IgG concentrations in women were apparently higher. Therefore, both smoking status and gender were regarded as very significant sources of variation in the interpretation of IgG values. However, the percentage of women who were cigarette smokers was much lower than the percentage of men. Given such a gender difference in IgG, the differences observed in IgG concentrations according to smoking status may simply reflect the underlying gender bias in the subgroup of smokers.

To clarify this point, reference values can be sub-grouped both by gender and smoking status. The result is shown in Figure 4. The left half of the Figure indicates that, among males, IgG is apparently lower in smokers. The same is true for females as shown on the right half of the Figure.

Furthermore, when we look at the Figure by focusing on subgroups of smokers (the first and the third subgroups), no gender difference is observed. The same is true for non-smokers (the second and fourth sub-groups). Thus, there is no gender difference. This Figure clearly demonstrates that the gender related difference seen in Figure 3B was not real, but was simply due to gender bias in smoking status. Therefore, we can conclude that the serum concentration of IgG is associated with smoking status, but not with gender. We call this the ''confounding'' phenomenon: the difference between men and women was spurious and caused by neglecting the status of smoking.

In this regard, it is important to note that any study to derive the RI is an observational study. In such studies, there are often many related sources of variation and, thus, univariate comparison of subgroups is often meaningless (4). To this point, consider further that smokers tend to drink more. Therefore, drinking status may also affect any conclusions drawn regarding smoking status. To prove these associations, it is necessary to resort to multi-way stratification of reference values, but as the groups studied are subdivided progressively to account for potential confounding variables, the number of reference observations assigned to each subgroup will become progressively smaller, impeding reliable conclusions to be drawn.

Nested ANOVA

The problem of confounding variables associated with univariate analysis can be overcome using a nested analysis of variance (ANOVA) (5) which allow simultaneous comparison of two or more sources of variation. A nested ANOVA separates the magnitude of variations attributable to each factor and expresses it in terms of the SD or coefficient of variation (CV). Figure 5 illustrates the application of a three-level nested ANOVA to judge the importance of city, gender, and age as sources of variation in a multi-center study conducted in Asia (6). Please note that the nested ANOVA requires categorical data for the analysis. In the example, gender and city were categorical, but age was a continuous variable. Therefore, age was recoded into decades as 20, 30, 40, and 50 before the analysis.

As a very simple numerical example, a two-level nested ANOVA is shown in Table 1A. The model data consists of test results of specimens from three individuals to derive the pure component of between-individual SD (SD_{pBI}). Each individual is sampled on three separate occasions to derive the pure component of the within-individual SD (SD_{pWI}). Furthermore, each specimen is measured twice on separate

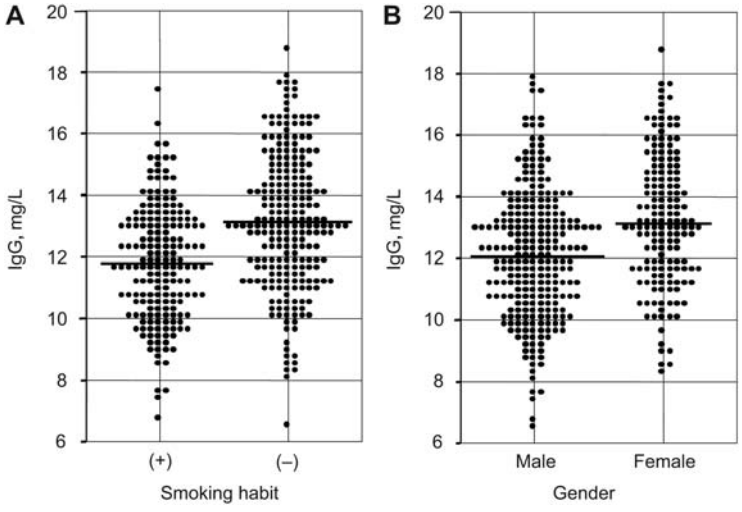

Figure 3 IgG associated with the smoking habit, gender, or both?
Test results of serum IgG obtained from 420 healthy subjects in a study of reference interval were subgrouped according to either smoking status (A) or gender (B). Both factors are significantly associated with the concentration of IgG.

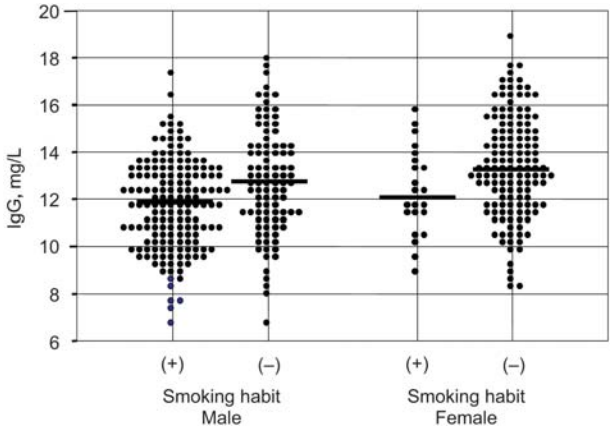

Figure 4 Two-way stratification to reveal independent association of smoking and gender with IgG.
The IgG concentration is apparently lower in smokers for both males and females, while no appreciable difference in IgG is observed between genders either among smokers or among non-smokers.

days to derive analytical SD (SD_A) which is denoted as residual variation and corresponds to the analytical precision.

The ANOVA tables are shown in Table 1B. The fourth column in the lower Table 1B reads $SD_{pBI}=9.663$, $SD_{pWI}=5.525$, and $SD_A=0.943$.

In the study of RI, we generally just measure each individual once, and we do not measure replicates of each specimen. Therefore, we cannot derive SD_{pWI} or SD_A, but we can derive SD_{BI}. Theoretically, the mean $\pm 1.96 \times SD_{BI}$ gives the RI. As a matter of fact, SD_{BI} obtained in such a study includes the components of variation due to SD_{pWI} and SD_A. Therefore, SD_{BI} can be expressed as follows.

$$SD_{BI}=\sqrt{SD_{pBI}^2+SD_{pWI}^2+SD_A^2}$$

In the numerical example shown in Table 2, the experiment is meant to evaluate how the magnitude of within individual SD or analytical SD influences SD_{BI} or the RI.

These theoretical considerations all assume an underlying Gaussian distribution of test results. In the real world, the distribution is often skewed. Therefore, such ANOVA computations should only be undertaken after transforming the values to obtain a Gaussian distribution. The RLs (LL^T and UL^T) in the transformed scale are computed. Then, they are reverse transformed to LL and UL on the original scale. An approximation of ''SD'' can be obtained as (UL–LL)/(1.96×2) (6).

Higher levels of nesting in nested ANOVA, which is not shown in conventional textbooks (5), are provided in advanced statistical packages, such as SAS® and R.

Multiple regression analysis

Multiple regression analysis provides the most powerful approach for evaluation of multiple potential sources (parameters) of variation simultaneously. It automatically adjusts for confounding influences of other parameters simply by including them concurrently in the regression model. The relative significance of each parameter can be predicted from the significance level of the regression coefficient. However, the magnitude of variation attributable to the significant parameter cannot be expressed as a SD or CV as with the nested ANOVA. Despite this limitation, multiple linear regression analysis can be useful for identifying potential parameters to be included in a nested ANOVA.

It is important to note that the analytical result of multiple linear regression analysis is greatly influenced by the presence of extreme values. Therefore, normalization of each variable by use of a power transformation, described below, is recommended to obtain optimal results. When using qualitative variables, such as gender, city, or blood type, dummy variables need to be introduced to judge their association with the test results of interest. There is no limitation on the number of variables that can be included in a multiple linear regression analysis. However, inclusion of too many variables in the regression model often hampers the reproducibility of the model.

For the analysis of data illustrated in Figures 3 and 4, the following equation allows us to judge the relative importance of different sources of variation in predicting the serum concentration of IgG.

$$IgG=b_0+b_1(\text{Gender})+b_2(\text{Smoking})+b_3(\text{Drinking})+b_4(\text{Age})+\cdots$$

Shown in Table 2B are the serial results of performing a stepwise multiple regression analysis on the sources of variations of serum IgG, using the dataset in Table 2A. The table illustrates the step by step changes in the regression parameters and their statistical significance with each stepwise addition of explanatory parameters.

In this analysis, dummy variables were introduced for gender (male = 0, female = 1), drinking status (Drk) and smoking status (Smk). The top table is the result with just gender in the model. It shows highly significant association with IgG. However, in the middle table, the result with Smk added to the model shows that the association of gender with IgG is reduced appreciably. Further addition of parameters Drk and age resulted in a non-significant association of gender with IgG. However, Smk remains significant in the presence of other competing factors, indicating that Smk is more directly associated with IgG. It is noteworthy that the association of many mutually related factors with the target variables can be examined easily without performing a multi-way stratification of the data as shown in Figure 4.

Criteria for partitioning reference values

Once a notable source of variation, such as gender has been identified using the statistical methods described above, it is important to determine the relative magnitude of variation associated with this source in order to evaluate whether it is necessary to develop separate RIs partitioned by this source. Several methods have been proposed for evaluating the need

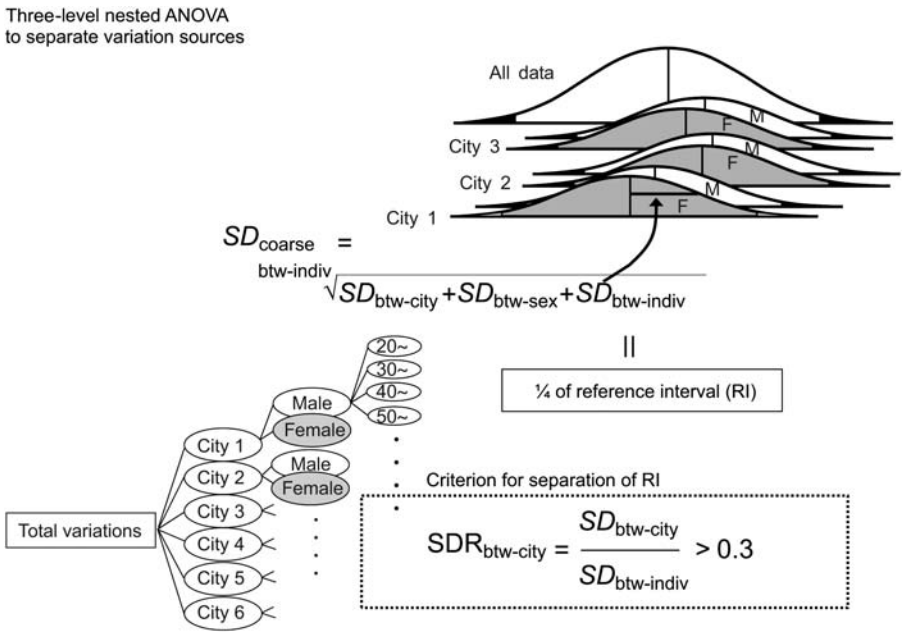

Figure 5 The concept of three-level nested ANOVA.
The illustration was made by supposing a dataset obtained in a multi-city study of reference interval (RI). The data are categorized by three factors: city, gender and age. The analysis separates sources of variations (SD) into three components. The relative magnitude of each SD is expressed as its ratio to between-individual SD, which corresponds to residual SD or 1/4 of the width of RI.

for partitioning by computing the ''effect size'' of given source of variation.

Harris-Boyd method

The Harris-Boyd method (1, 7) is applicable to the case when there are two subgroups. This method is based on principles of statistical tests for the differences between two sets of data, but it differs from the conventional test for two sample means, whose power is affected by the sample size. Rather, the method evaluates the practical significance of the difference between means by making adaptations for differing sample sizes. When this approach is used to judge the necessity for gender-specific RIs, it appears to yield recommen-

Table 1 A simple example for two-level nested ANOVA.

Table 1A

Subj	Day	Result
1	1	23
1	1	25
1	2	25
1	2	24
1	3	27
1	3	25
2	1	28
2	1	28
2	2	35
2	2	34
2	3	39
2	3	40
3	1	52
3	1	50
3	2	48
3	2	48
3	3	37
3	3	36

Table 1B Two-level nested ANOVA.

Source var	Sum of sqr	df	Variance	F-value	p-Value
Btw indiv var	1244.3	2	622.17	10.04	0.0122
Wtn indiv var	371.7	6	61.94	69.69	0.0000
Residual	8.0	9	0.89		
Total var	1624	17			

Analysis of variance component (VC).

Source var	VC	VC, %	SD	CV (VC)
Btw indiv var	93.370	74.824	9.663	27.874
Wtn indiv var	30.530	24.464	5.525	15.938
Residual	0.890	0.712	0.943	2.720
Sum of var	124.79			
Grand mean	34.67			

Source var, source of variations; Btw indiv var, between-individual variations; Wtn indiv var, within individual variations; Total var, total variations; df, degree of freedom.

Table 2 An example for multiple regression analysis.

Table 2A

IgG	Age	Sex	Drk	Smk
13.09	54	0	0	0
17.94	48	0	1	0
9.86	44	0	1	0
11.48	42	0	1	0
16.24	52	0	1	0
15.73	59	0	0	0
11.31	52	0	1	0
17.34	60	0	1	0
6.66	53	0	1	0
15.05	38	0	1	0
12.58	55	0	1	0
16.49	51	0	1	0
10.37	35	0	1	0
12.41	41	0	1	0
13.35	36	0	1	0
10.37	41	0	1	0
15.13	52	0	1	0
11.48	52	0	0	0
11.22	49	0	1	0
8.84	40	0	1	0
11.56	46	0	1	0
12.92	58	0	0	0
14.37	56	0	1	0
13.18	50	0	1	0

n=420.

Table 2B

Multiple regression analysis.

Target variable: IgG, g/L; n=420.

Order	Var name	b	SE(b)	t-Value	df	p-Value
0		12.128	0.131			
1	Sex	1.143	0.210	5.448	418	0.000

Multiple corr coeff R=0.2575. Adjusted R^2=0.0641.

Order	Var name	b	SE(b)	t-Value	df	p-Value
0		12.772	0.194			
1	Sex	0.637	0.235	2.710	417	0.007
2	Smk	−1.022	0.231	4.429	417	0.000

Multiple corr coeff R=0.3290. Adjusted R^2=0.1040.

Order	Var name	b	SE(b)	t-Value	df	p-Value
0		11.733	0.661			
1	Sex	0.449	0.251	1.791	415	0.074
2	Smk	−0.944	0.231	4.082	415	0.000
3	Drk	−0.436	0.240	1.820	415	0.069
4	Age	0.028	0.012	2.284	415	0.023

Multiple corr coeff R=0.3561. Adjusted R^2=0.1184.

The original data shown in Figures 3 and 4, which were obtained in a multi-center study of reference intervals in Japan. Var name, name of explanatory variable; b, regression coefficient; SE(b), standard error of b; df, degree of freedom for t-value; Multiple corr coeff, multiple correlation coefficient; Drk, status of drinking alcohol; Smk, cigarette smoking status.

dations that agree with those that have been used traditionally in setting the RIs for analytes with known gender-related differences (Figure 6). To aid in understanding this method, it is helpful to know that if we assume that the data size of the two subgroups (n_1, n_2) are the same and the SD of the two subgroups (s_1, s_2) are the same and equal to 's'. Then, using the factor of 3.0 described in the original description of this method (7), the critical value of the difference between the means of subgroups is 0.387s (if the 3.0 factor is adjusted to 5.0, as was later suggested by Harris-Boyd (1), this critical value becomes 0.645s). This implies that, independent of the sample size, the critical difference between subgroups is approximately 38% of the size of SD_{BI}: that is 's', or about one fourth of the RI obtained after separating the two groups. Flaws in the Harris-Boyd method have been pointed out by Lahti et al. (8), which has led to development of the Lahti method.

Lahti method

The second criterion, proposed by Lahti (9), is based on the percentages of reference values in each subgroup lying outside the ULs and the LLs of the RI derived without partitioning. It focuses on the statistically unstable peripheral part of the distribution. The drawbacks of this method are its limited applicability to the study of RIs without having a large sample size, and unsuitability to the case in the presence of multiple subgroups, some of which may have small sample sizes.

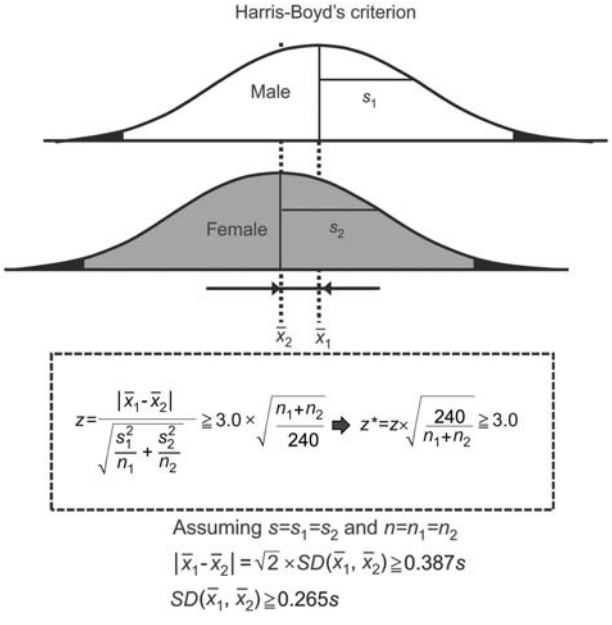

Figure 6 The concept of the Harris-Boyd's method.

Fraser method

The third criterion, proposed by Fraser et al. (10), is based on the magnitude of biological variation expressed as the standard deviation (SD_B), which includes both SD_{BI} and SD_{pWI} components. The latter includes the component of analytical variation, and is thus equivalent to S_{BI} described above. If the SD accounting for between-subgroup variation is $0.375 \times SD_B$, separate RIs are to be considered.

Ichihara method

The Ichihara method depends on a two- or three-level nested ANOVA (6). It utilizes the information on SD attributable to each source of variation. It computes the SD ratio (SDR), which is the SD of a given factor divided by the SD_{BI}, or SD due to between-individual variation shown in Figure 5. This method requires that the underlying data being analyzed be transformed to yield a Gaussian distribution (6).

One advantage of this method is that can be applied to the situation where there are more than two subgroups categorized by the factor. Another merit is that the method is suited to situations in which there is a confounding influence of other factors as illustrated in Figures 3 and 4. This ability to account for confounding variables allows the method to yield judgment regarding the necessity of partitioning after adjustment for the confounding variables, something that none of the other methods offer. The critical value can be arbitrarily set according to the policy in deriving RIs. In the previous survey for reference values (6) with a data size of 580, we adopted $SDR > 0.4$ as a cut-off value to judge the necessity of partitioning. But in the recent survey with a much larger data set, we found it more appropriate to use $SDR = 0.3$ as the cut-off value to match the judgment to the customary setting of gender-specific RI (see the description below regarding Table 3). This method can be compared with that of the Harris-Boyd method when dealing with the problem of two subgroups. Since the SD for between individual corresponds to 's' in the Harris-Boyd method and the SD of two values, a and b, is $|a-b|/\sqrt{2}$, the 0.3 limit for SDR gives a criterion that is a little more conservative in judging the need for partitioning compared to the one given by the Harris-Boyd method, which corresponds to 0.265s. However, the Ichihara method is a little stricter than that of Frazer since the latter criterion corresponds to the 0.375 limit for SDR.

In Table 3, judgments using the Harris-Boyd and Ichihara methods are compared with respect to between-gender difference using a dataset obtained from the recent multicenter study. The judgments using the cut-off values of $Z^* > 3.0$ with the Harris-Boyd method and $SDR = 0.3$ with the Ichihara method match very closely.

Techniques for exclusion of extreme values

Univariate approach

It is important to note that the presence of an extreme value can be judged only by properly assuming the distribution pattern of reference values at hand. Therefore, if the non-parametric method is adopted in deriving RI, it is not possible to judge extreme values in a strict sense. However, some ''non-parametric'' robust methods are available that have some limitations. One of the most commonly used methods is the Tukey method described by Horn et al. (11). This method uses the inter-quartile range (IQR) as a yardstick. IQR is the interval between the 1st quartile (Q1) and the 3rd quartile (Q3), that is, $IQR = Q3 - Q1$. When a given data point in the distribution is above $Q3 + 1.5 \times IQR$, or below $Q1 - 1.5 \times IQR$, it is considered an outlier. If the distribution is Gaussian, the Tukey criterion theoretically leads to deletion of 0.7% of data in the tails. Solberg and Lahti found the Tukey method to be relatively insensitive for the detection of outliers (12).

The Dixon test (13) is another ''non-parametric'' method that focuses on the most extreme value (X_n) and subtracts from it the next most extreme value (X_{n-1}) as $D = X_n - X_{n-1}$, and then divides it by the range of the values, R (where $R = X_n - X_1$), as D/R. If D/R exceeds 0.3, X_n is regarded as an outlier. The Dixon test is rather conservative and cannot deal with situations where there are multiple extreme values.

In cases, where an underlying Gaussian distribution is assumed, perhaps after power transformation of reference values, the common practice is to delete values outside the $\text{mean} \pm k \times SD$ range. 'k' is usually set to 3.0 or larger. It is important to note that if one sets 'k' at a small value, say 2.0, and applies the exclusion procedure repetitively, the data size is reduced progressively, thus distorting the original distribution. Therefore, more robust ways are required for judgment including the Smirnov method (14), Healy method (15), or an iterative method proposed by Ichihara and Kawai (16). These work well under the assumption of a Gaussian distribution, and thus are used in external quality control surveys (15, 16) for which the test results tend to follow a Gaussian distribution since the values have been obtained by repetitive measurements of the same specimens.

On the contrary, the data used for determination of reference values are generally not assumed to be Gaussian, and the dilemma is that the method required for Gaussian transformation is easily influenced by the presence of extreme values. Therefore, the CLSI guideline recommends a conservative policy not to delete any values since there is no generally applicable method of outlier detection.

Multivariate approach

It is not always a good policy to judge any value simply by its location in the distribution. Rather, the judgment should be made according to other information related to the test values concerned. Such an approach has been used by Kratzsch et al. who followed The National Academy of Clinical Biochemistry (NACB) recommendations for determining the appropriate reference population to use in developing RIs for thyroid stimulating hormone (TSH). In this approach, individuals were excluded if they had a family history of thyroid disease, positive autoantibodies to thyroid peroxidase (TPO) and/or thyroglobulin (Tg), increased free triiodothy-

Table 3 Comparison of two partitioning criteria in judging gender-related differences using a dataset from a multicenter study.

Item	Unit	Male			Female			Harris-Boyd method			Nested ANOVA		
		n	Mean	SD	n	Mean	SD	Z	Constant	Z*	SDR-sex	SDR-city	SDR-age
TP	g/L	1437	71.4	3.9	1871	71.4	4.1	0.22	0.269	0.1	0.00	0.36	0.23
Alb	g/L	1437	43.2	2.6	1871	41.8	2.5	15.97	0.269	4.3	0.40	0.00	0.45
UN	mmol/L	1437	4.8	1.1	1871	4.3	1.1	14.35	0.269	3.9	0.33	0.24	0.35
UA	μmol/L	1437	349	65	1871	246	50	50.04	0.269	13.5	1.30	0.19	0.13
CRE	μmol/L	1437	77	10	1871	56	7	69.98	0.269	18.9	1.83	0.24	0.07
Na	mEq/L	1436	142.9	1.5	1871	141.9	1.6	17.78	0.269	4.8	0.44	0.00	0.30
K	mEq/L	1436	4.21	0.29	1871	4.14	0.29	6.98	0.269	1.9	0.08	0.23	0.20
Cl	mEq/L	1436	103.9	1.8	1871	104.5	1.7	9.33	0.269	2.5	0.23	0.00	0.18
Ca	mmol/L	1436	2.36	0.07	1871	2.32	0.07	14.15	0.269	3.8	0.35	0.00	0.34
IP	mmol/L	1437	1.18	0.16	1871	1.26	0.15	14.31	0.269	3.9	0.36	0.00	0.34
Glu	mmol/L	1437	4.92	0.52	1871	4.72	0.47	11.65	0.269	3.1	0.27	0.16	0.39
TCho	mmol/L	1437	4.94	0.85	1871	4.96	0.89	0.36	0.269	0.1	0.00	0.00	0.51
TG	mmol/L	1437	1.23	0.66	1871	0.85	0.44	18.93	0.269	5.1	0.56	0.27	0.36
HDL-C	mmol/L	1437	1.24	0.32	1870	1.55	0.39	24.73	0.269	6.7	0.55	0.57	0.00
LDL-C	mmol/L	1437	3.06	0.76	1871	2.84	0.76	7.88	0.269	2.1	0.16	0.00	0.46
AST	U/L	1437	24.2	6.0	1871	21.2	5.4	15.02	0.269	4.0	0.41	0.12	0.33
ALT	U/L	1437	28.6	12.5	1871	20.1	8.2	22.32	0.269	6.0	0.68	0.12	0.30
LD	U/L	1437	188	27	1871	182	28	6.18	0.269	1.7	0.12	0.08	0.34
ALP	U/L	1437	64	16	1871	55	15	16.48	0.269	4.4	0.43	0.15	0.35
GGT	U/L	1435	36.0	20.6	1857	23.4	12.1	20.71	0.270	5.6	0.69	0.01	0.34
CK	U/L	1437	136	78	1871	83	38	23.81	0.269	6.4	0.82	0.11	0.18
AMY	U/L	1437	81	26	1871	86	27	6.18	0.269	1.7	0.15	0.11	0.12
CRP	g/L	1411	0.97	1.77	1831	0.77	1.56	3.22	0.272	0.9	0.00	0.42	0.15
IgG	g/L	1437	12.1	2.3	1870	12.9	2.3	11.00	0.269	3.0	0.15	0.38	0.06
IgA	g/L	1437	2.46	0.72	1871	2.54	0.73	3.15	0.269	0.8	0.00	0.27	0.07
IgM	g/L	1436	1.05	0.40	1871	1.54	0.64	26.48	0.269	7.1	0.71	0.17	0.34
C3	mg/L	1435	1061	181	1871	1020	180	6.48	0.269	1.7	0.00	0.47	0.22
C4	mg/L	1436	216	60	1871	207	62	4.49	0.269	1.2	0.00	0.41	0.25

The judgments of the Harris-Boyd method (Z*) and Ichihara method (SDR) based on the three-level nested ANOVA were compared with respect to the necessity of subgrouping by gender: Z*>3.0 and SDR>0.3, respectively. The values of SDR for between-city and between-age differences are also shown for the latter method. The gray colored cells indicate those analytes which were judged as requiring separation by gender in deriving the RI. Z* represents Z statistics adjusted for data sizes of subgroups. Z* value exceeding 3.0 generally indicates significant difference between two subgroups.

ronine and/or free thyroxine, or sonographically assessed abnormalities of the thyroid (17). The latent abnormal value exclusion (LAVE) method is an example of multivariate-based judgment of extreme values. This method was first used in a study by Ichihara and Kawai to derive RIs for serum proteins in a Japanese population (18). It has been used extensively in later studies of RI in Japan (6, 19–22). The method has been developed to exclude abnormal results hidden within the reference values. However, it does not judge how extreme a given test value might be in isolation.

Instead, this method looks at other concurrently measured test results. The method is a type of iterative approach for derivation of multiple RIs simultaneously, in which no exclusion of values is made in the initial computation of RIs. The algorithm then uses those initial values of RIs to judge the abnormality of each individual's record by counting the number of abnormal results in tests other than the one for which the RI is being determined (Figure 7). This algorithm is potentially useful for data mining RIs from data accumulated in a routine screening setting, which may contain a certain percentage of abnormal individuals. Grossi et al. (23) used a modification of this approach with an additional criterion that cases could be excluded only when the concurrent test abnormalities occurred in tests that were *significantly correlated* with the results of the test for which the RI was being computed.

Actually, the most challenging issue in the derivation of RIs in the healthy adult population is how to discern those who are affected by common conditions, such as metabolic syndrome or related disorders. To set strict exclusion criteria to those analytes known to be influenced by such disorders [triglyceride, uric acid, alanine aminotransferase (ALT), γ-glutamyltransferase (GGT), etc.] results in unrealistic reduction in the data size and truncation of the reference distribution (narrow RIs). Meanwhile, setting lenient criteria leads to wider RIs disparate from the clinical decision limits generally set by consensus for those analytes. In such a case, the LAVE method generally gives intermediate RIs as long as mutually related analytes are measured together. The advantage of the LAVE method is that truncation of the reference distribution does not occur since the decision to exclude a record from any individual is made only by the results of other analytes that have been measured concurrently. However, when there is no association among test results, RIs are not influenced at all, but the data size remaining after the final iteration is reduced to a certain extent depending on the associations between the other analytes. This reduction in data size is an apparent disadvantage when

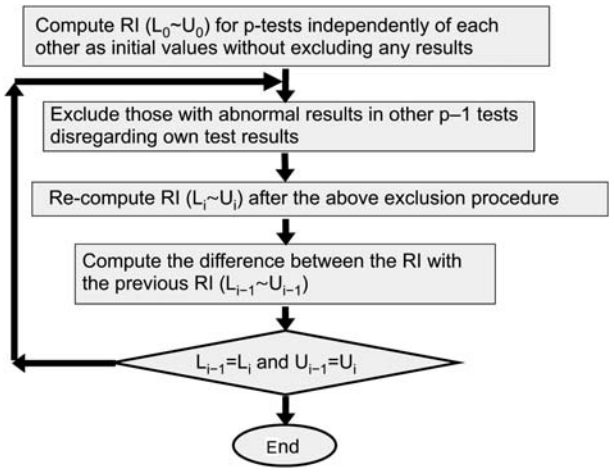

Figure 7 The computation flow of latent abnormal values exclusion (LAVE) method.

the original data size is not large. Therefore, judicious use of the LAVE method is necessary by planning which group of analytes to include in the simultaneous derivation of RIs, and which analytes to be used for judging latently abnormal records. In general, for analytes whose test results may be used as exclusion criteria, we recommend those which are closely associated with common disorders, such as metabolic syndrome, diabetes, or anemia. They include glucose, triglyceride, low density lipoprotein-cholesterol (LDL-C), low density lipoprotein-cholesterol (HDL-C), uric acid, ALT, aspartate aminotransferase (AST), GGT, hemoglobin (Hb), etc. Meanwhile, the analytes for which abnormal results are infrequent among healthy individuals are not to be included in the list: Na, K, Cl, creatinine, alkaline phosphatase (ALP), urea, platelet count, etc.

Computation of reference interval

Parametric method

The parametric method makes use of the theory that when the distribution is Gaussian, the CI for any probability can be computed by use of the mean and SD of the distribution. RI, which is the 95% CI of the distribution, is thus obtained as the mean $\pm 1.96 \times$ SD. Therefore, it is mandatory to ensure that the distribution is Gaussian before computing RI parametrically.

Since almost all distributions of laboratory test results are non-Gaussian, it is essential to convert these to a Gaussian distribution. The power transformation described below is particularly useful for this purpose.

Power transformation The most flexible way to achieve Gaussian transformation is by use of the modified Box-Cox formula. The Box-Cox transformation was initially recommended by the IFCC Expert Committee on RIs as an approach for transformation of reference population data to a Gaussian distribution (24). The modified Box-Cox formula was developed by Ichihara and Kawai (18) and takes the form shown in Figure 8, where 'p' and 'a' represent power and the origin of the transformation.

The original Box-Cox equation does not include 'a'. However, this constant is essential for successful transformation when the distribution arises far away from zero as illustrated in Figure 9.

In Figure 9, the original reference distribution has a typical log-normal pattern. The reference values start a little above five. Therefore, unless a certain amount is subtracted from each observation to account for displacement of the origin, the log transformation fails as in cases [2] and [4]. Meanwhile, the same log-transform with subtraction of five, as in case [6], results in a successful Gaussian transformation. It also can be noted that when the power transformation was done less vigorously, by specifying $p = 0.333$ (or cubic square root), the transformation was less complete under the setting of the same origin. However, it is notable that [2] and [3], or [4] and [5] resulted in comparable distribution patterns judging from skewness and kurtosis.

This is because if the origin 'a' is set to an optimal position close to the lowest values of the distribution, the power 'p' can be set to a larger value compared with the case of transformation where 'a' is neglected. Thus, the values of 'a' and 'p' are closely associated. Such an association requires a special technique to be used to derive an optimal maximum likelihood estimate (MLE) of the two parameters. Several such techniques are potentially available.

The performance of the Box-Cox transformation, with or without the constant to correct for displacement of the origin, was compared using the dataset (NHANES III) provided in the textbook on RI by Horn and Pesce (2). Table 4 shows the analytical results for men. The adequacy of transformation to a Gaussian form by the original Box-Cox and the modified Box-Cox method are tested for each analyte. The fitting to the Gaussian transformation can be judged by the skewness and kurtosis of the distribution where $|\text{skewness}|$ should < 0.15 and $2.7 < \text{kurtosis} < 3.3$ for an adequate transformation. The darker gray colors for the columns of Skew and Kurt indicate non-fitting to the Gaussian distribution. It is apparent that the original Box-Cox model was not successful in converting most of the variables to a Gaussian distribution.

Probability paper method: is it a valid method? In the past, as a modification of the parametric method, normal probability paper (x-axis-test value; y-axis = cumulative frequency from 0 to 1) has been used extensively to derive RI by graphically truncating the non-linear part of the curve (cumulative frequency curve). This method was first proposed by Pryce (25) and supported by Hoffman (26). The following assumptions are made in proposing this method: [1] the reference distribution is roughly categorized into normal or log-normal, [2] non-linearity on the probability paper is attributable to the presence of an abnormal group of individuals skewing the distribution, [3] extrapolation of the linear segment in the left half of the curve gives ULs of the RI at the intersection with the horizontal red line of $y = 0.975$,

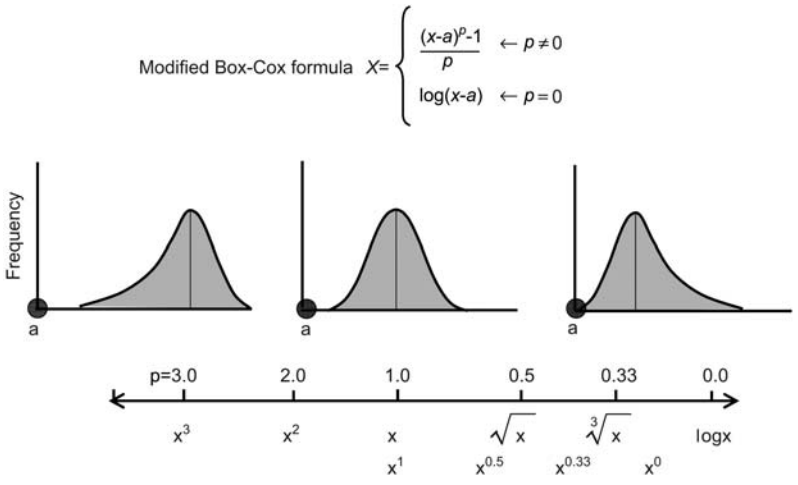

Figure 8 The concept of modifed Box-Cox formula.
The power 'p' can take any value except zero. In cases of p = 0.0, the equation is switched to the logarithmic formula. The origin of the transformation 'a' is generally set at a mean–4.0SD of the distribution to get optimal transformation of the data into the Gaussian distribution.

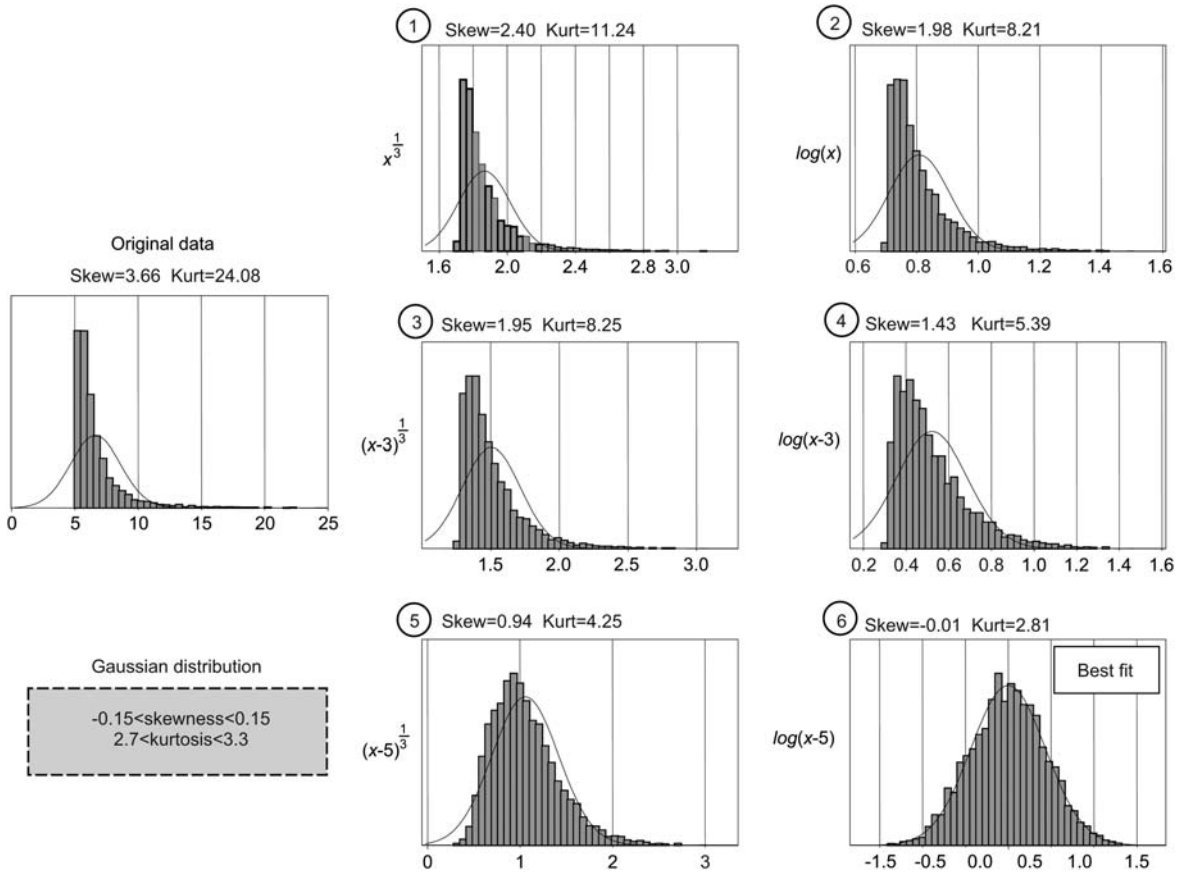

Figure 9 Importance of the origin in successful power transformation.
The original distribution shown in the left is supposed to be a logarithmic normal distribution. Five different formulae were used for power transformation. The results are influenced both by the power 'p' (either 0.33 or 0.0) and the origin 'a'. The goodness of fit to Gaussian transformation in terms of skewness and kurtosis is shown at the top of each graph.

thus removing the abnormal group. Figure 10 demonstrates the feasibility of these claims by simulation, introducing ''abnormal groups'' at variable locations and with varying sample sizes. The main Gaussian distribution was set as a constant. The straight line in each panel indicates the theoretical line corresponding to the main peak. It is apparent that a hand

Table 4 Performance of the modified Box-Cox formula for Gaussian transformation.

Male		Box-Cox			Modified Box-Cox				LL	Me	UL
Item	n	p-Value	Skew	Kurt	p-Value	a	Skew	Kurt			
WBC	3955	−0.20	−0.16	3.26	0.47	2.6	−0.11	3.05	3.9	6.8	11.4
Lym	3905	0.10	−0.02	3.39	0.04	−0.4	−0.07	2.88	1.2	2.2	3.7
Mon	3848	0.40	0.03	5.32	0.58	0.0	−0.03	3.20	0.1	0.4	0.8
Gran	3851	0.00	−0.14	3.41	0.37	0.6	−0.11	3.02	1.9	4.0	7.6
RBC	3911	0.90	−0.22	4.70	0.79	3.9	−0.03	3.01	4.4	5.0	5.7
Hb	4029	1.00	−0.58	5.13	0.90	115.8	−0.16	2.88	133.5	151.2	170.0
Ht	4027	1.00	−0.47	5.21	0.91	0.4	−0.07	2.86	0.4	0.4	0.5
MCV	3872	1.00	−0.38	5.51	1.08	72.1	0.00	3.08	81.5	89.7	97.5
MCH	3875	1.00	−0.68	5.96	1.30	23.7	−0.02	3.07	27.1	30.4	33.3
MCHC	3920	1.00	−0.02	5.11	1.14	305.3	0.00	2.99	322.6	338.5	353.5
PLT	3955	0.50	0.17	5.00	0.27	24.3	−0.09	3.11	165.4	258.0	388.7
Na	3880	0.60	−0.13	5.44	1.12	133.0	0.06	3.10	137.6	141.7	145.6
K	3913	0.20	0.10	3.76	0.72	3.2	−0.01	2.89	3.5	4.0	4.6
Cl	3926	1.00	−0.39	4.90	1.08	92.5	−0.05	2.92	98.9	104.6	110.2
CO2	3965	−0.50	−0.05	3.67	0.36	18.5	−0.26	3.57	22.5	28.1	37.2
Ca	3928	−0.50	−0.06	6.88	0.99	2.0	0.03	2.99	2.2	2.3	2.5
IP	3931	0.40	0.04	5.03	0.83	0.6	−0.10	3.01	0.8	1.1	1.4
UA	3914	0.60	0.15	3.62	0.58	131.5	−0.12	3.09	230.9	350.0	505.8
Glu	3870	−1.00	1.09	9.50	0.60	3.7	0.07	3.23	4.2	5.1	6.4
UN	3973				0.56	1.2	−0.12	3.12	2.7	5.0	8.0
TBil	3870	0.00	−3.66	75.54	0.00	−1.0	−0.20	4.94	4.7	10.5	22.6
CRE	3957	−0.50	0.33	12.74	0.05	47.4	−0.43	3.60	79.8	100.2	132.6
AST	4018	−0.70	0.21	4.44	0.03	5.3	−0.16	5.35	13.1	21.5	38.5
ALT	4002	−0.20	0.20	4.16	0.03	3.2	−0.20	3.72	7.8	18.2	50.2
GGT	3147	−0.50	−0.08	3.21	0.03	7.8	−0.23	4.11	11.1	26.4	102.5
LD	3955	0.00	0.30	5.44	0.53	88.1	−0.06	3.06	108.2	153.2	221.7
ALP	3904	0.00	0.46	6.37	0.13	−5.0	−0.09	3.19	48.6	81.8	131.7
TP	3926	0.10	−0.12	4.62	0.87	59.7	−0.07	2.77	66.6	74.5	83.0
Alb	3973	1.00	−0.42	6.73	0.93	31.2	0.02	2.94	37.2	43.3	49.7
Posm	3008	1.00	0.18	5.37	0.92	258.0	−0.18	2.68	268.6	279.6	291.1
UCRE	3895				0.61	−0.3	0.01	2.88	2.4	13.8	31.3

The dataset (NHANES III) in Horn's book (2) was used to compare the performance of the modified Box-Cox formula with that of original Box-Cox formula used by Horn. Goodness of fit to Gaussian distribution was shown after power transformation by either methods in terms of skewness and kurtosis of the distribution. With a Gaussian distribution, the skewness should be close to zero and kurtosis approximately 3.0. The denser the cell color, the more bias from the Gaussian form.

drawn line might not coincide with the blue line and could easily skew towards the actual curve depicted from the entire distribution. In addition, the method did not appear to give good estimates of RIs when it was tried in practice (27, 28).

The inherent problem of this methodology is its basic assumption that laboratory results in healthy individuals follow a normal (or log-normal) distribution. As is shown in Table 4, the power 'p' predicted using the modified Box-Cox equation is often different than either p = 1.0 (normal) or p = 0.0 (log-normal). Therefore, when the distribution cannot be well categorized as normal or log-normal, truncation of the non-linear portion of data on normal probability paper results in a falsely narrow RI.

Non-parametric method

Simple percentile estimation When the reference values show a peculiar skewed distribution which cannot be made Gaussian, even by use of the power transformation, or when there are many values below the detection limit of the assay, the non-parametric method is the method of choice for computing RI.

This method estimates the 2.5 percentile for the LL and the 97.5 percentile for the UL.

There are many ways to compute percentile estimates. The following three approaches are the most common. The percentile **p** of the **r**-th value **x** (**x**[**r**]) with the data size of **n** is formulated. As a simple numerical example, a list of percentiles corresponding to five values (**n** = 5: **x**[**1**], **x**[**2**], **x**[**3**], **x**[**4**], **x**[**5**]) are shown within the parenthesis.

$$\text{Method 1: } p = \frac{r-0.5}{n} \times 100 \quad (10 \quad 30 \quad 50 \quad 70 \quad 90)$$

$$\text{Method 2: } p = \frac{r}{n+1} \times 100 \quad (16.7 \quad 33.3 \quad 50 \quad 66.7 \quad 83.3)$$

$$\text{Method 3: } p = \frac{r-1}{n-1} \times 100 \quad (0 \quad 25 \quad 50 \quad 75 \quad 100)$$

It should be noted that the p estimates are symmetrically allocated on both sides of the central data **x**[**3**], but their ranges differ between them. It is interpreted that Method 2 assumes the presence of data on both ends of the distribution by leaving space open for the 0 and 100 percentiles, while

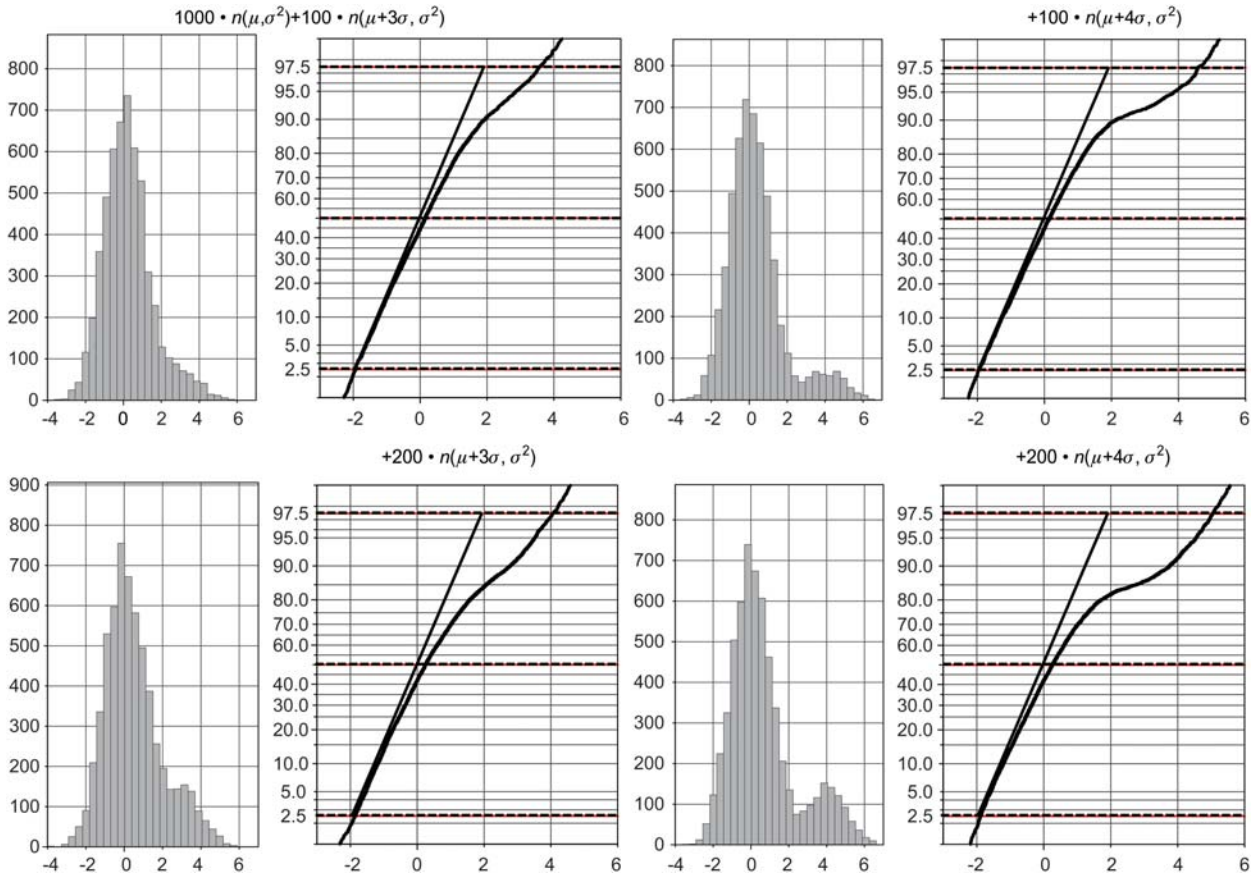

Figure 10 Validity of probability paper method.
As a simulation, a sample assuming to have a Gaussian distribution of mean = μ and SD = σ, designated as n (μ, σ) with sample size of 1000 was generated. It was merged with another sample with a Gaussian distribution generated with a variable center and size, but with the same SD = σ. The merged distribution is shown both as a histogram and plotted on probability paper. The blue line on the latter corresponds to the theoretical line corresponding to the main sample. In the probability paper method for derivation of RI, a linear segment of the curve is derived manually. This figure points out the apparent difficulty of obtaining an optimal linear segment by the manual procedure.

Method 3 does not assume the presence of other data by occupying the 0 and 100 percentiles with real data. However, Method 1 gives the percentile estimate to be placed between them. Therefore, in cases with a small n, the medium assignment of the percentile by Method 1 appears more reasonable, although the results of the three methods become identical with increments in n (29).

For non-parametric derivation of RI, estimation of the 2.5 and 97.5 percentile by any of the three methods is required. Estimation of the 2.5 percentile is illustrated as follows.

1) Find rank r corresponding to 2.5 percentile

Method 1: $\mathbf{r = 2.5 \times \frac{n}{100} + 0.5}$

Method 2: $\mathbf{r = 2.5 \times \frac{n+1}{100}}$

Method 3: $\mathbf{r = 2.5 \times \frac{n-1}{100} + 1}$

If the computed **r** is an integer with no fractional value, then the **r**th value, **x**[**r**], corresponds to the 2.5 percentile. Otherwise, interpolation is required to obtain 2.5 percentile.

2) Estimate percentiles p_1 and p_2 closest to 2.5 percentile

The **r** is rounded down to an integer and values **x** at ranks **r** and **r+1** are determined. The estimates of p_1 and p_2, corresponding to the values **x**[**r**] and **x**[**r+1**], are given as follows:

Method 1: $\mathbf{p_1 = \frac{r-0.5}{n} \times 100 \quad p_2 = \frac{r+1-0.5}{n} \times 100}$

Method 2: $\mathbf{p_1 = \frac{r}{n+1} \times 100 \quad p_2 = \frac{r+1}{n+1} \times 100}$

Method 3: $\mathbf{p_1 = \frac{r-1}{n-1} \times 100 \quad p_2 = \frac{r+1-1}{n-1} \times 100}$

3) Derive the value corresponding to 2.5 percentile

Using the following equation of linear interpolation, the estimate of the value of the 2.5 percentile is obtained by use of p_1, p_2, **x**[**r**], and **x**[**r+1**].

$$2.5 \text{ percentile} = x[r] + \frac{2.5 - p_1}{p_2 - p_1} \times (x[r+1] - x[r])$$

The 97.5 percentile is similarly estimated by replacing 2.5 with 97.5 in the above formulae.

More precise percentile estimation As described in the first part of this review, simple non-parametric estimation of RI is generally inferior in precision to parametric methods, especially with a small sample size. Various approaches to increase the precision of the estimate have been proposed in the book by Harris-Boyd (1) (pages 35–39). They include [1] the weighted average of all the observed percentiles using a filter function called the kernel or [2] application of the so-called bootstrap or re-sampling method. In the latter procedure, the complete set of n observations is sampled repeatedly, each resample built up by n sequential random selections with replacement. Thus, any resample may include the same observation more than once, or not at all. The RI (LL, UL) is derived from each resampled dataset. After repetition of this process, usually more than 500 times, the average of LL and UL estimates gives a smoothed non-parametric estimate of the RI. The advantage of this method over the kernel based method is that it provides both the smoothed estimate and its standard error.

Conclusions

Various statistical methods and computational techniques must be considered at each step of any study performed for deriving RIs. This review has emphasized the following points.

The CI of RLs is strongly tied to the underlying sample size. To obtain reliable RIs, it is essential to conduct a multicenter study with recruitment of a large number of reference individuals.

To determine factors that influence the test values in the reference population, the test results should be analyzed using a multivariate method to avoid confounding influences of other factors. Multiple regression analysis is recommended for this purpose. To judge the necessity for partitioning RI, various criteria are proposed. When there are multiple categories in a given factor, nested ANOVA may be the method of choice with its additional capability for simultaneous comparison of multiple factors.

In computing RIs, the parametric method often requires the use of power transformation to make the distribution Gaussian. The Box-Cox transformation formula is often used for this purpose, but of note, it does not function well without adjustment for the origin of transformation. The CIs of RLs derived by the parametric method with transformation are narrower than those for the non-parametric method, but the margin of difference is rather small owing to the additional error introduced in the estimation of the parameters, the power factor and the origin when using the parametric method. The merit of using the parametric method is its capability of identifying and potentially excluding extreme values during computation of RIs.

Acknowledgments

We are grateful to Dr. Ferruccio Ceriotti of Diagnostica e Ricerca San Raffaele, Milan, and Dr. Joseph Henny, Laboratoire de Biologie Clinique, Centre de Médicine Préventive, Vandoeuvre-lès-Nancy, France, for their invaluable comments reviewing this article.

Conflict of interest statement

Authors' conflict of interest disclosure: The authors stated that there are no conflicts of interest regarding the publication of this article.
Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.

References

1. Harris EK, Boyd JC. Statistical basis of reference values in laboratory medicine. New York: Marcel Dekker, 1995.
2. Horn PS, Pesce AJ. Reference intervals. A user's guide. Washington, DC: AACC Press, 2005.
3. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline, 3rd ed. CLSI document C28-A3. Clinical and Laboratory Standards Institute, Wayne, PA, USA, 2008.
4. Ichihara K, Kawai T. Determination of reference intervals for 13 plasma proteins based on IFCC international reference preparation (CRM470) and NCCLS proposed guideline (C28-P, 1992): a strategy for partitioning reference individuals with validation based on multivariate analysis. J Clin Lab Anal 1997; 11:117–24.
5. Sokal RR, Rohlf FJ. Biometry, 3rd ed. New York: W.H. Frieman, 1995:272–320.
6. Ichihara K, Itoh Y, Lam CW, Poon PM, Kim JH, Kyono H, et al. Sources of variation of commonly measured serum analytes among 6 Asian cities and consideration of common reference intervals. Clin Chem 2008;54:356–65.
7. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. Clin Chem 1990;36: 265–70.
8. Lahti A, Petersen PH, Boyd JC, Rustad P, Laake P, Solberg HE. Partitioning of non-Gaussian-distributed biochemical reference data into subgroups. Clin Chem 2004;50:891–900.
9. Lahti A. Are the common reference intervals truly common? Case studies on stratifying biochemical reference data by countries using two partitioning methods. Scand J Clin Lab Invest 2004;64:407–30.
10. Fraser CG, Hyltoft Petersen P, Libeer JC, Ricos C. Proposals for setting generally applicable quality goals solefly based on biology. Ann Clin Biochem 1997;24:8–12.
11. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and nonhealthy individuals on reference interval estimation. Clin Chem 2001;47:2137–45.
12. Solberg HE, Lahti A. Detection of outliers in reference distributions: performance of Horn's algorithm. Clin Chem 2005;51: 2326–32.
13. Dixon WJ. Processing data for outliers. Biometrics 1953;9: 74–89.
14. Smirnov NV. On the estimation of the maximum term in a

series of observations [in Russian]. Dokl Akad Nauk SSSR, 1941;33:346–9.
15. Healy MJ. Outliers in clinical chemistry quality-control schemes. Clin Chem 1979;25:675–7.
16. Ichihara K, Kawai T. An iterative method for improved estimation of the mean of peer-group distributions in proficiency testing. Clin Chem Lab Med 2005;43:412–21.
17. Kratzsch J, Fiedler GM, Leichtle A, Brügel M, Buchbinder S, Otto L, et al. New reference intervals for thyrotropin and thyroid hormones based on National Academy of Clinical Biochemistry criteria and regular ultrasonography of the thyroid. Clin Chem 2005;51:1480–6.
18. Ichihara K, Kawai T. Determination of reference intervals for 13 plasma proteins based on IFCC international reference preparation (CRM470) and NCCLS proposed guideline (C28-P,1992): trial to select reference individuals by results of screening tests and application of maximal likelihood method. J Clin Lab Anal 1996;10:110–7.
19. Ichihara K, Itoh Y, Min WK, Yap SF, Lam CW, Kong XT, et al. Diagnostic and epidemiological implications of regional differences in serum concentrations of proteins observed in six Asian cities. Clin Chem Lab Med 2004;42:800–9.
20. Ichihara K, Saito K, Itoh Y. Sources of variation and reference intervals for serum cystatin C in a healthy Japanese adult population. Clin Chem Lab Med 2007;45:1232–6.
21. Tamechika Y, Iwatani Y, Tohyama K, Ichihara K. Insufficient filling of vacuum tubes as a cause of microhemolysis and elevated serum lactate dehydrogenase levels. Use of a data-mining technique in evaluation of questionable laboratory test results. Clin Chem Lab Med 2006;44:657–61.
22. Matsubara A, Ichihara K, Fukutani S. Determination of reference intervals for 26 commonly measured biochemical analytes with consideration of long-term within-individual variation. Clin Chem Lab Med 2008;46:691–8.
23. Grossi E, Colombo R, Cavuto S, Franzini C. The REALAB project: a new method for the formulation of reference intervals based on current data. Clin Chem 2005;51:1232–40.
24. 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;25:645–56.
25. Pryce JD. Level of haemoglobin in whole blood and red blood-cells, and proposed convention for defining normality. Lancet 1960;2:333–6.
26. Hoffman RG. Statistics in the practice of medicine. J Am Med Assoc 1963;185:864–73.
27. Amador E, Hsi BP. Indirect methods for estimating the normal range. Am J Clin Pathol 1969;52:538–46.
28. Elveback LR, Guiller CL, Keating FR Jr. Health, normality, and the ghost of Gauss. J Am Med Assoc 1970;211:69–75.
29. Linnet K. Non-parametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 2000;46:867–9.