Accred Qual Assur (2013) 18:51–60
DOI 10.1007/s00769-012-0948-5

PRACTITIONER'S REPORT

The comparison of parametric and nonparametric bootstrap methods for reference interval computation in small sample size groups

Abdurrahman Coskun · Elvan Ceyhan · Tamer C. Inal · Mustafa Serteser · Ibrahim Unsal

Received: 11 September 2012 / Accepted: 4 December 2012 / Published online: 18 December 2012

Abstract According to the IFCC, to determine the population-based reference interval (RI) of a test, 120 reference individuals are required. However, for some age groups such as newborns and preterm babies, it is difficult to obtain enough reference individuals. In this study, we consider both parametric and nonparametric bootstrap methods for estimating RIs and the associated confidence intervals (CIs) in small sample size groups. We used data from four different tests [glucose, creatinine, blood urea nitrogen (BUN), and triglycerides], each in 120 individuals, to calculate the RIs and the associated CIs using nonparametric and parametric approaches. Also for each test, we selected small groups ($m = 20, 30, \ldots, 120$) from among the 120 individuals and applied parametric and nonparametric bootstrap methods. The glucose and creatinine data were normally distributed, and the parametric bootstrap method provided more precise RIs (i.e., the associated CIs were narrower). In contrast, the BUN and triglyceride data were not normally distributed, and the nonparametric bootstrap method provided better results. With the bootstrap methods, the RIs and CIs of small groups were similar to those of the 120 subjects required for the nonparametric method, with a slight loss of precision. For original data with normal or close to normal distribution, the parametric bootstrap approach should be used, instead of nonparametric methods. For original data that deviate significantly from a normal distribution, the nonparametric bootstrap should be applied. Using the bootstrap methods, fewer samples are required for computing RIs, with only a slightly increased uncertainty around the end points.

Keywords Box-Cox transformation · Confidence interval · Nonparametric bootstrap · Parametric bootstrap · Percentile · Reference interval

Abbreviations

RI	Reference interval
CI	Confidence interval
URIL	Upper reference interval limit
LRIL	Lower reference interval limit
BUN	Blood Urea Nitrogen
CLSI	Clinical and Laboratory Standards Institute
IFCC	International Federation of Clinical Chemistry and Laboratory Medicine
ISO	International Standardization for Organization

Introduction

Clinical laboratory test results are obtained from thousands of patients, and the interpretation of these results is a comparative decision-making process that requires a reference interval (RI) for each test [1]. We need reference values of all tests performed in clinical laboratories. Reference value is obtained by measurement of a particular type of quantity on a reference individual [2]. Reference values are not identical for all individuals; they have a dispersion termed the reference distribution. In general, RI is accepted as the central 95 % interval of the reference

A. Coskun (✉) · T. C. Inal · M. Serteser · I. Unsal
Department of Biochemistry, School of Medicine, Acibadem University, Gulsuyu, Maltepe, Istanbul, Turkey
e-mail: coskun2002@gmail.com

E. Ceyhan
Department of Mathematics, College of Science, Koç University, Istanbul, Turkey

distribution bounded by the 2.5 and 97.5 percentiles. Since the concept of a reference value was first introduced by Grasbeck and Saris [3], reference values have been the subject of many scientific studies and are widely accepted by medical professionals [4]. Scientific organizations such as the Clinical and Laboratory Standards Institute (CLSI) and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) recommend that reference values to be determined in individual laboratories by measuring analyte levels in a group of reference individuals who are healthy and represent the population served by the laboratory [5, 6]. According to these recommendations, each laboratory would determine RIs for all of the tests performed in the laboratory using its own methods and reagents. In addition to clinical laboratories, manufacturers of lab kits are obliged to provide reference limits in package inserts, according to directive 98/79/CE of the European Community.

The current approach to establishing RIs is based on the recommendations of the IFCC and CLSI [5, 7], which require the analysis of 120 individuals to determine the population-based RI for a test. However, 120 individuals may not be adequate in certain situations, for example, a reference population with a skewed distribution [8]. Individuals of a population are usually different from each other, and consequently, the reference population would be expected to be heterogeneous. To address these issues, we use a partitioning method to obtain homogenous data, which requires a greater number of reference individuals. Additionally, for some age groups, such as newborns and preterm babies, it is difficult to obtain adequate data for determining RIs. For these reasons, despite their logical approach and strong scientific background, the recommendations of the CLSI and IFCC have not been followed worldwide. Although many scientists accept this approach theoretically, only a few laboratories have been able to follow the recommendations in practice.

To overcome the sample size problem, alternative methods such as robust and bootstrap methods have been developed [9, 10]. Robust methods are too complicated and not practical for clinical laboratories. Among the bootstrap methods, only nonparametric methods have been used in limited studies.

In the present study, we consider nonparametric bootstrap methods for the estimation of RIs and the associated CIs for the end points of RIs in small sample size groups ($20 \leq m \leq 120$). We also touch base on data transformations to normality, which is necessary prior to parametric estimations of RIs. Furthermore, for the first time, we propose parametric bootstrap methods applied to the transformed data to construct RIs. We also provided a guide for practical computation of RIs and the associated CIs for the end points using bootstrap methods.

Methods

The data sets used in this study were obtained from the Acibadem Labmed Clinical Laboratories (Istanbul, Turkey), which is the first clinical laboratory in Turkey to be awarded the ISO 15189 accreditation standard. It provides multidisciplinary clinical chemistry, hematology, immunology, microbiology, and virology laboratory services in a central laboratory for 14 different hospital laboratories and five outpatient clinical laboratories.

For the present study, we selected four different tests performed in our laboratory: glucose, BUN, creatinine, and triglycerides. To evaluate the RIs, we selected 120 individuals for each test from our laboratory database. All data were selected from the check-up unit of Acibadem Hospital. For each test, we randomly selected small groups (m individuals, $m = 20, 30, \ldots, 120$) from the $n = 120$ individuals and applied nonparametric and parametric bootstrap methods to compute the RIs and associated CIs for the end points. The CIs are constructed for the lower and upper end points of the RIs, to determine the precision of the RIs and hence are referred to as "the associated CIs" for the RI end points. To distinguish the sample sizes used for bootstrap sampling and the generally suggested sample size for RI interval calculation, we use m for the former and n for the latter.

To determine the optimal number of replications, we also determined the RIs and associated CIs based on nonparametric bootstrap methods with 999 or 9999 replications.

Nonparametric approach to RI estimation

Let $X_1, X_2, \ldots, X_m$ be m measurements or observations in a random sample from a population of interest with cumulative distribution function F with mean μ and standard deviation σ. Denote the order statistics as $X_1 < X_2 < \ldots < X_m$. A $100(1-\alpha)$ % RI means that $100(\alpha/2)$ % of values are below the left end of RI and $100(\alpha/2)$ % of the values are above the right end of the RI.

Computation of a $100(1-\alpha)$ % RI by the nonparametric approach requires to find the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the random sample (which would be estimates of the given percentiles of the distribution F). For example, for a 95 % RI, the standard limits in the clinical chemistry are 2.5th and 97.5th percentiles of the data set. The simplest way to find these percentiles in the nonparametric approach is finding the values with the appropriate rank. The pth percentile of a sample is the value whose rank is $p(n+1)$. If $p(n+1)$ is not an integer, we perform a linear interpolation between the two order statistics whose ranks surround $p(n+1)$. For example, for $m = 999$, the 2.5th percentile is the order statistics whose

rank is 25 (i.e., 25th value when ordered from smallest to largest), and for $m = 99$, the 2.5th percentile is the value whose rank is 2.5, so linear interpolation would imply taking the average of second and third smallest values. The RIs estimated in this way are also referred to as "quantile RI estimates."

These estimates of RIs are distribution free or nonparametric (i.e., do not require the data to be of a certain distribution), and perform quite well (i.e., robust) for large samples. However, they are usually less efficient and have large variance [11]. Robust estimators for smaller samples are also proposed in the literature. For example, Horn et al. [9] propose a method that performs better than transformation to normality in estimating more accurate reference limits and also Harrell and Davis [12] method performs well for skewed data with $20 \leq m \leq 60$. As an alternative to many such robust methods, bootstrapping is also proposed to estimate RIs.

As for the CIs for the end points of the RIs, we apply the CI estimation for a quantile based on binomial distribution and normal approximation to binomial [13]. In particular, for continuous F, let $F(x_p)$, that is, let x_p be the $(100p)$th percentile of F. Then, $100(1-\alpha)$ % CI for x_p is $(X_{(i)}, X_{(j)})$ with $B(j-1, n, p) - B(i-1, n, p)$ being as close as possible to $1-\alpha$, where $B(k,n,p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{(n-i)}$ (which is the cumulative distribution function for a binomial distribution with n trials and success probability being p). When such i and j do not yield the exact confidence, we apply a linear interpolation in i in the right end and in j in the left end of the RI. For large n, a normal approximation can also be used. For example, let $B(k-1, n, p) = \Phi(z)$ where $z = (k-1+0.5-np)/\sqrt{(np(1-p))}$ and $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. So for a specified confidence level $1-\alpha$, setting $z = z_{1-\alpha/2}$ we obtain $i = 0.5 + np + z_{1-\alpha/2}\sqrt{np(1-p)}$ and $j = 0.5 + np - z_{1-\alpha/2}\sqrt{np(1-p)}$. If i and j are rounded to nearest integers, then $(X_{(i)}, X_{(j)})$ would be the desired CI (or if i and j are not integers, one can perform a linear interpolation to find the appropriate order statistics, which is the practice we adopt in this article).

Parametric approach to RI estimation

The most common parametric approach to RI estimation is based on assuming F to be the normal distribution with mean μ and standard deviation σ, denoted as $N(\mu, \sigma)$. Then, a $100(1-\alpha)$ % RI has end points $\hat{\mu} \pm z_{1-\alpha/2}\hat{\sigma}$ where $\hat{\mu}$ and $\hat{\sigma}$ are the sample estimates of the mean and standard deviation. Although this interval gives the exact desired level for RIs, it is not realistic in practice [14], since positive skewness is common in clinical and biological measurements. Many transformation techniques (to normality) are recommended to overcome this difficulty of deviation from normality [11].

The CIs for the end points of the RIs can be computed based on the binomial approach or normal approximation to binomial approach.

Nonparametric bootstrap methods for RI estimation

The nonparametric bootstrap approach to RI computations involves sampling m values with replacement from the original data set of size m, thereby generating a "bootstrap" sample. For each bootstrap sample, we found the empirical percentiles, with linear interpolation as necessary. For example, if $m = 100$, and we want to find the 90 % empirical percentiles, we take the 5th and 95th (smallest) values in the sample. We repeated this procedure R times to obtain R bootstrap samples. For each bootstrap sample, we compute the lower and upper limits (i.e., percentiles) as estimates of the reference limits so that we obtain R lower and upper limits as estimates of the reference limits from the bootstrap samples. The mean of the R bootstrap percentiles provided the bootstrap estimates of the RIs. At each bootstrap replication, it is possible to use other nonparametric methods for the RI calculation and then take the average of the obtained RI estimates.

There are many methods for constructing CIs based on nonparametric bootstrapping, including the basic bootstrap CI, studentized CI, percentile CI, adjusted percentile CI (BCa), ABC CI, and normal approximation CI [15, 16]. Here, we used only basic, percentile, and normal approximation bootstrap CIs for the end points of the RIs, as these are commonly used and are simpler than the other methods.

In particular, let T be the estimated quantity from the original sample (T is $100(1-\alpha/2)$th or $100(\alpha/2)$th percentile in our case). That is, if we have 1000 estimated values, and RIs with 95 % would have 25th and 975th values as upper and lower limits, respectively. Let T_r^* be the corresponding estimate for bootstrap sample r. Also, let $T_{(k)}^*$ is the kth value of T_r^* values. Then, the basic CI is estimated as

$$\left(2T - T^*_{((R+1)(1-\alpha))}, 2T - T^*_{((R+1)\alpha)}\right). \quad (1)$$

The percentile CI is estimated as

$$\left(T^*_{((R+1)\alpha)}, T^*_{((R+1)(1-\alpha))}\right) \quad (2)$$

and normal approximation CI is estimated as

$$\left(T + \sqrt{V_R} z_{1-\alpha}, T - \sqrt{V_R} z_{1-\alpha}\right) \quad (3)$$

where $\upsilon_R = \frac{1}{R-1}\sum_{r=1}^{R}\left(T_r^* - \overline{T}^*\right)^2$ with $\overline{T}^* = \frac{1}{R}\sum_{r=1}^{R} T_r^*$ and z_α is the $100(1-\alpha)th$ percentile of the standard normal distribution. There is also a bias correction possible for normal approximation CI [16].

Table 1 RIs and associated CIs obtained using the nonparametric method, for glucose, BUN, creatinine, and triglyceride tests performed in our laboratory

Test ($n = 120$)	95 % RI	90 % CIs for reference limits	
		Binomial distribution	Normal approximation
Glucose (mmol/L)	4.33–6.33	(4.12, 4.44) (6.06, 6.83)	(4.12, 4.44) (6.11, 6.37)
BUN (mmol/L)	2.43–6.85	(2.32, 3.00) (6.57, 7.68)	(2.32, 2.96) (6.64, 7.68)
Creatinine (μmol/L)	40.7–92.8	(36.2, 47.7) (91.1, 98.1)	(36.2, 47.7) (91.9, 98.1)
Triglyceride (mmol/L)	0.49–2.76	(0.41, 0.60) (2.65, 2.97)	(0.41, 0.60) (2.66, 2.97)

RIs are at the 95 % level; CIs for the ends of RIs are at the 90 % level

Parametric bootstrap methods for RI estimation

In general, in the parametric bootstrap method for a data set from a distribution F which is specified by the parameter θ, each bootstrap sample is generated from F with an estimate of θ from the original sample [16]. In particular, if $F = N(\mu, \sigma)$, then each bootstrap sample is generated from a $N(\hat{\mu}, \hat{\sigma})$ distribution, where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and standard deviation, respectively, based on the original sample. To estimate the RI by the parametric bootstrap method, the required percentile estimates were obtained from each generated sample, and RIs are the means of these sample percentiles. Associated CIs for the end points of the RIs were obtained as in the nonparametric case, that is, basic, percentile, and normal approximation methods for CIs.

Results

Table 1 presents the RIs and associated CIs for the end points for the four data sets (with $n = 120$) we considered. The RIs were constructed using the nonparametric approach, while the CIs were constructed based on binomial distribution and normal approximation to the

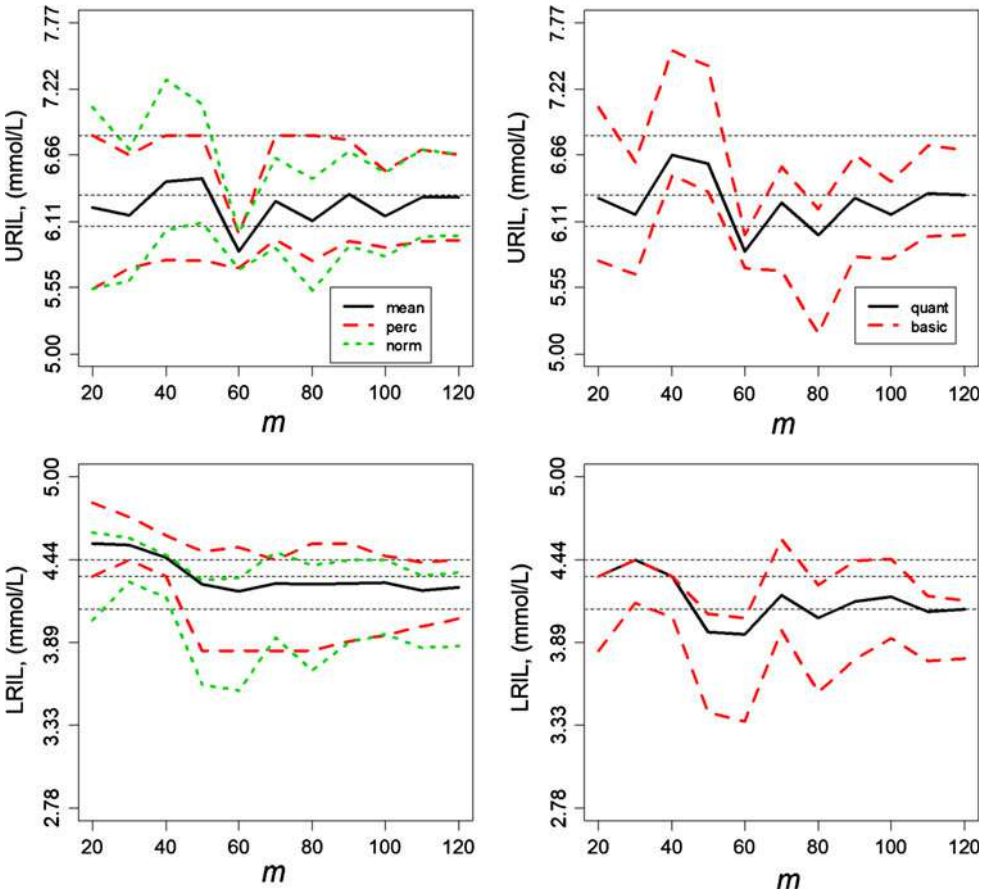

Fig. 1 RI estimates based on the nonparametric bootstrap method for measurements of glucose concentrations with sample sizes $m = 20, 30, \ldots, 120$. For each sample size, 999 bootstrap samples were generated. Upper reference limits are plotted in the *top row*, and lower limits are plotted in the *bottom row*. *Horizontal dashed lines* are 95 % RIs, and the 90 % CIs around the reference limits based on nonparametric estimation of RIs with 120 observations. "Mean" is the average of the reference limits from bootstrap samples; "perc" is the percentile bootstrap CI estimate, "norm" is the normal approximation bootstrap CI, "basic" is the basic bootstrap CI for reference limits, and "quant" is the quantile estimate of the reference limit based on the bootstrap samples. The estimated RIs and the associated CIs for $m \leq 120$ are joined by straight lines for better visualization. (*URIL* upper reference interval limit, *LRIL* lower reference interval limit)

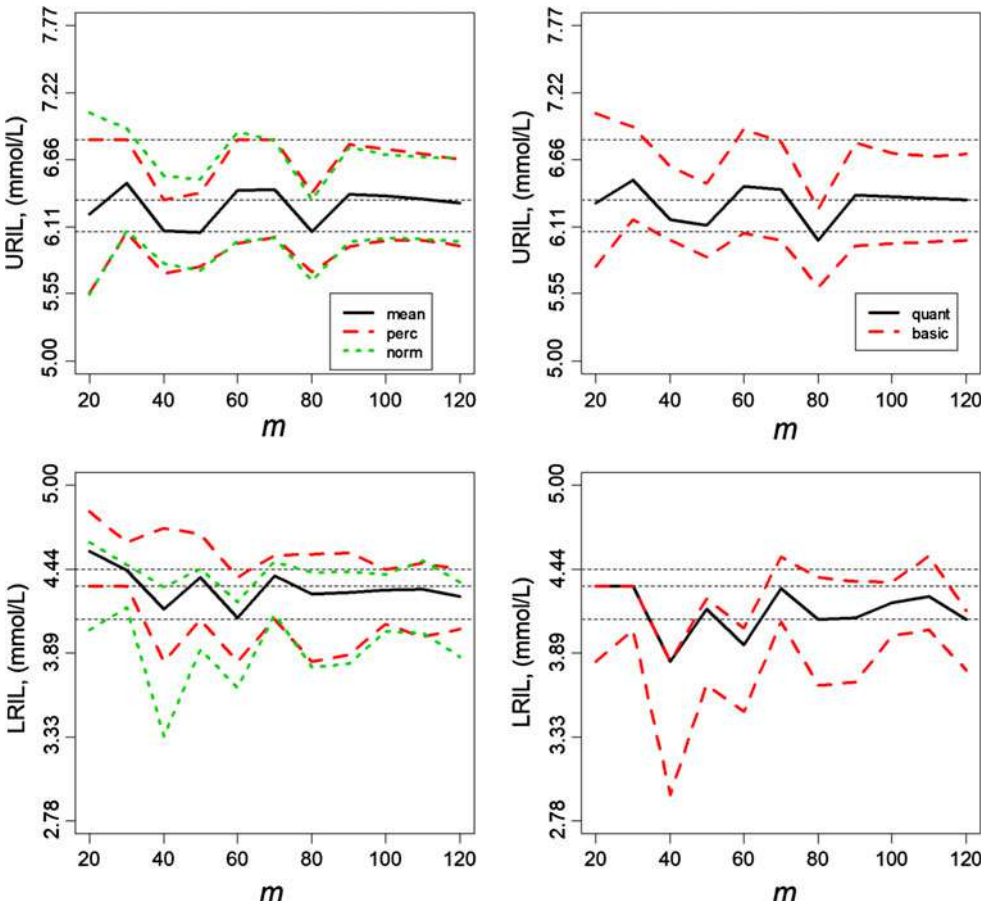

Fig. 2 RI estimates based on the nonparametric bootstrap method for measurements of glucose concentration with sample sizes $m = 20$, 30,…, 120. For each sample size, 9999 bootstrap samples were generated. Figure and legend descriptions are as in Fig. 1

binomial. Because the sample size n is large ($n = 120$), the CIs are very similar between the two methods (together with linear interpolation).

For the glucose test, we constructed the RIs and associated CIs based on nonparametric bootstrap methods with $R = 999$ and $R = 9999$ replications. Figure 1 presents the nonparametric bootstrap RIs and associated basic, percentile, and normal approximation CIs with $R = 999$ replications, with different sample sizes ($m = 20$, 30,…, 120) as well as with $n = 120$ for comparative purposes. The bootstrap RIs and associated CIs with $R = 9999$ replications are presented in Fig. 2. Notice that a tenfold increase in the number of bootstrap samples caused only a slight decrease in the bandwidth of the CIs and did not improve the RIs (i.e., they were not narrower than those with $R = 999$). As a similar trend was observed for the other three tests, the bootstrap RIs and associated CIs for the other tests with $R = 9999$ replications are not presented. The nonparametric bootstrap RIs and associated CIs for the other three tests (BUN, creatinine, and triglycerides) with $R = 999$ are presented in Fig. 3, where the basic bootstrap CIs for the end points are omitted. The trends are similar to those of the glucose test shown in Fig. 1.

In the parametric construction of the RIs for the glucose data, we first test for normality ($p = 0.2741$ based on the Anderson–Darling test). Although this is not significant evidence against normality, we also performed a Box-Cox transformation on the glucose data to further get it close to the normal distribution. For the other data sets, we followed a similar procedure. For the triglyceride data, we also performed an IFCC transformation to correct for the kurtosis.

Table 2 presents the RIs and associated CIs for the end points obtained using the parametric approach. The CIs based on the binomial distribution and normal approximation were very similar because of the large sample size. Compared with the nonparametric RIs in Table 1, the parametric RIs were similar for the glucose, BUN, and creatinine data. The parametric RIs for the triglyceride data based on the Box-Cox transformation were much wider than their nonparametric counterparts, such that the lower ends were very similar, but the upper ends were much higher. Parametric RIs for the triglyceride data based on the IFCC transformation were narrower than those with the Box-Cox transformation, but were still wider than the nonparametric counterparts, such that the lower ends were increased, but the upper ends were still higher. The CIs for

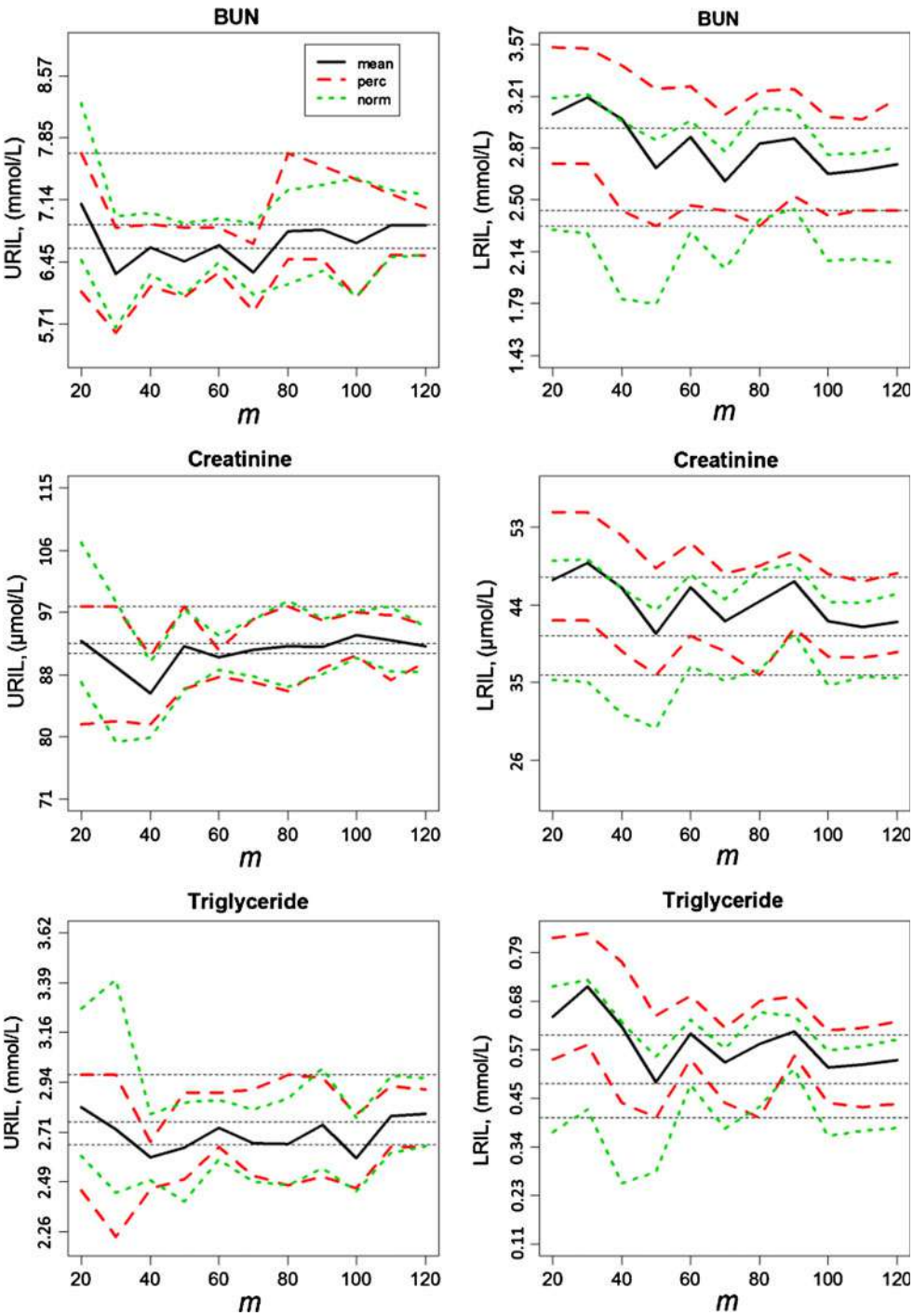

Fig. 3 RI estimates based on the nonparametric bootstrap method for measurements of BUN, creatinine, and triglyceride concentrations with sample sizes $m = 20$, 30,…, 120. For each sample size, 999 bootstrap samples were generated. Figure and legend descriptions are as in Fig. 1 (except basic CIs are not presented)

the end points are very similar between the nonparametric and parametric RIs, as they are computed using the same method.

The RIs and associated CIs for the end points based on the parametric bootstrap method for the glucose, BUN, and creatinine data with Box-Cox transformation are presented in Fig. 4. With the parametric bootstrap, the RIs for all $m \geq 20$ were narrower than those for the $n = 120$ original data points. We recommend that the basic bootstrap CI should be avoided, because it may miss the corresponding reference limit estimate, and a CI should at least contain the estimate of the parameter it is used for. Compared with a normal approximation CI, a percentile CI has a lower upper end and a higher lower end of the RI.

The RIs and associated CIs for the end points for the triglyceride data with Box-Cox and IFCC transformations based on the parametric bootstrap method are presented in Fig. 5. Again, with the parametric bootstrap, the RIs for all

Table 2 RIs and associated CIs for the end points of RIs obtained using the parametric method, for glucose, BUN, creatinine, and triglyceride tests performed in our laboratory

Test ($n = 120$)	95 % RI	90 % CIs for reference limits	
		Binomial distribution	Normal approximation
Parametric method with $n = 120$ (based on Box-Cox transformation)			
Glucose (mmol/L)	4.27–6.29	(3.83, 4.44) (6.06, 6.83)	(3.83, 4.44) (6.11, 6.83)
BUN (mmol/L)	2.75–7.10	(2.32, 3.00) (6.57, 7.68)	(2.32, 2.96) (6.64, 7.68)
Creatinine (μmol/L)	41.5–96.4	(36.2, 47.7) (91.1, 98.1)	(36.2, 47.7) (91.9, 98.1)
Triglyceride (mmol/L)	0.49–3.07	(0.41, 0.60) (2.65, 2.97)	(0.41, 0.60) (2.66, 2.97)
Parametric method with $n = 120$ (based on transformation recommended by IFCC)			
Triglyceride (mmol/L)	0.53–2.97	(0.41, 0.60) (2.65, 2.97)	(0.41, 0.60) (2.66, 2.97)

RIs are at the 95 % level; CIs for the ends of RIs are at the 90 % level

$m \geq 20$ with both transformations were narrower than those for the $n = 120$ original data points. However, percentile CIs are lower for both ends of the RIs compared to those of normal approximation CIs. With IFCC transformation, RIs are slightly narrower and so are the associated CIs compared to the ones with Box-Cox transformation.

The discrete values at integers are joined for better visualization in Figs. 1, 2, 3, 4, and 5. However, in Figs. 1, 2 and 3, samples of size m only point out to an increase in the sample size, and no clear trend occurs, so a linear or any other type of interpolation would not be meaningful. For example, it will not be possible to estimate a value at $m = 22.3$ in these figures. But in the parametric bootstrap figures (i.e., in Figs. 4, 5), there seems to be a continuous trend; hence, a value at $m = 22.3$ can be approximated by linear interpolation. Notice also that erratic fluctuations occur for nonparametric bootstrap RIs, not for parametric bootstrap RIs. Because in nonparametric bootstrap, we use random samples of size $m = 20$, 30,.., 120 measurements from the original sample of $n = 120$ measurements, as our original data sets from which bootstrap samples are taken. Hence, each sample of size m is different and may result in shifts of RI estimates, say from $m = 20$ to 30. On the other hand, parametric bootstrap RIs are based on sampling from the fitted normal distribution (after a transformation is employed); hence, bootstrap samples are generated from approximately normal distributions.

Discussion

Sample size is a significant issue in calculating the RI. Currently, nonparametric methods are widely used to calculate RIs. However, neither parametric nor nonparametric methods should be applied to limited sample sizes such as 20 or 40. To overcome this limitation, new statistical methods are required. In the present study, we evaluated both parametric and nonparametric bootstrap methods for the calculation of RIs and the associated CIs for the end points for small sample size groups in different tests. We demonstrated that both parametric and nonparametric bootstrap methods are powerful tools to calculate RIs and the associated CIs in small sample size groups.

In the bootstrap method, a set of data is randomly resampled, with replacement, multiple times (about 1000 and 10000 times), and statistical calculations such as RI and CI are performed using this large data collection. Although bootstrapping is a computer-based procedure, it was a time consuming and impractical method for determining the RI in clinical laboratories. However, most clinical laboratories now have improved computing power and are able to apply this method.

In the present study, using nonparametric and/or parametric bootstrap methods with fewer data, we showed that it is possible to obtain results that are close to the ones obtained from the nonparametric construction of RIs with $n = 120$ values. When the RIs and CIs for the glucose test were determined using the nonparametric bootstrap method, increasing the number of bootstrap samples tenfold ($R = 999$ to $R = 9999$ replications) caused only a slight reduction in the bandwidth of the CIs and did not improve the RIs (Figs. 1, 2). A similar trend was observed for the other tests. Hence, the smaller bootstrap sampling with $R = 999$ replications can be used, although there is only a small gain in computation time compared to $R = 9999$ replications. In comparison, Wright and Royston [11] recommended at least 500 bootstrap samples. The bootstrap-estimated CIs were as wide as or wider than the CIs based on the nonparametric estimation. That is, the bootstrap RIs had more variation than the nonparametric RI estimates. Thus, the bootstrap approach does not provide the same precision as the nonparametric method, but the RIs can be similar, with a little more uncertainty for the end points of the RIs.

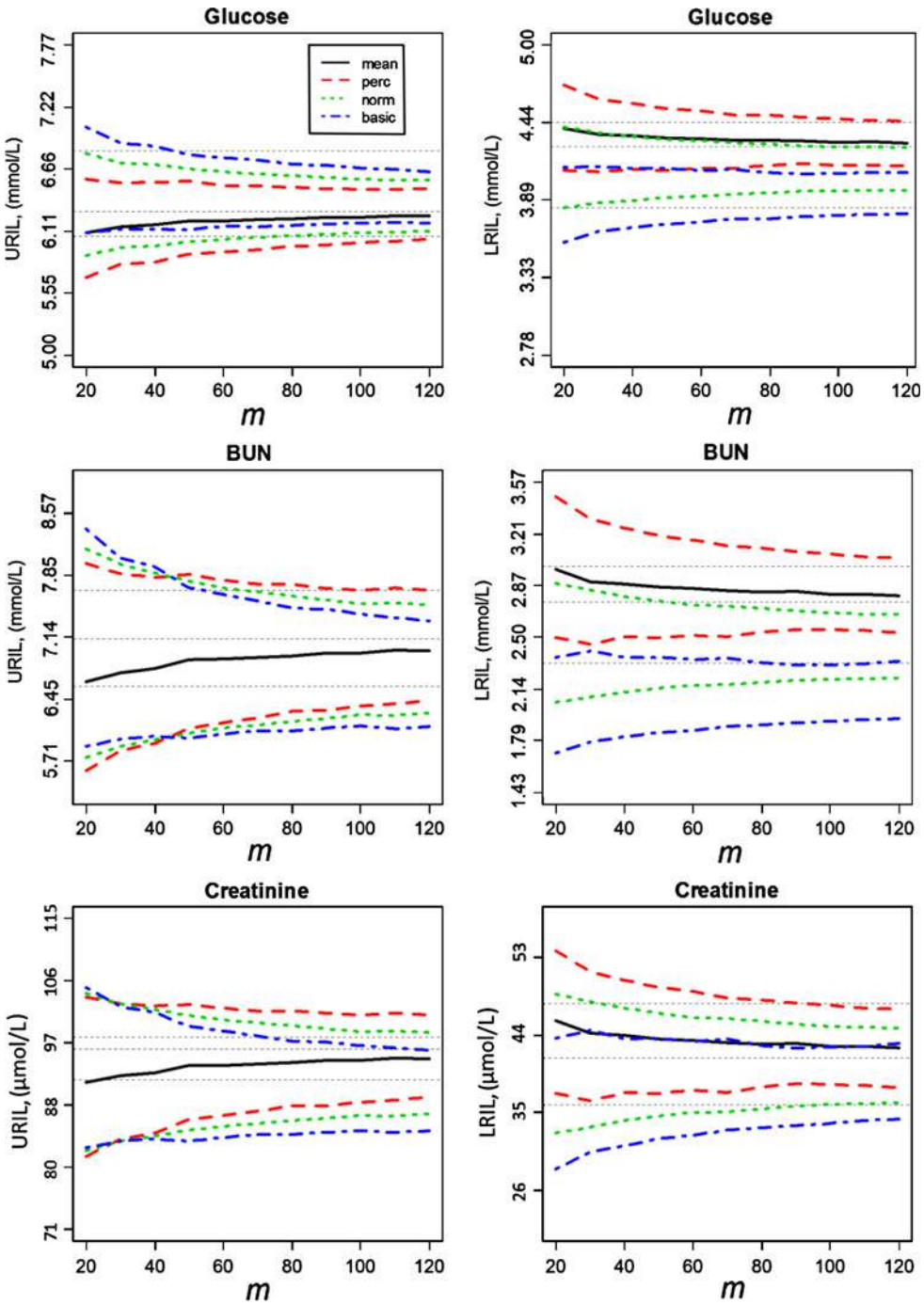

Fig. 4 RI estimates based on the parametric bootstrap method for measurements of glucose, BUN, and creatinine concentrations with Box-Cox transformation and with sample sizes $m = 20, 30, \ldots, 120$. For each sample size, 999 bootstrap samples were generated. *Horizontal dashed lines* are 95 % RIs and the 90 % CIs around the reference limits based on parametric estimation of RIs with 120 observations. Legend labeling is as in Fig. 1

The trends for the BUN, creatinine, and triglyceride tests were similar to the glucose test results (Fig. 3): The RIs tended to be narrower, and the percentile CIs were slightly lower at the upper end and substantially higher at the lower end of the RIs. For narrower RIs, the percentile bootstrap CIs are recommended when the nonparametric bootstrap method is employed to construct the RIs; otherwise, normal approximation CIs are preferred.

For the parametric bootstrap estimation of RIs and associated CIs for the end points, the transformation that renders the distribution closest to normality should be selected. The IFCC transformation [7] usually performs better in this regard, but requires more effort. As for the CIs, either percentile or normal approximation CIs can be used, depending on the data.

For the glucose and creatinine data, the parametric bootstrap method provided more precise RIs, as evidenced by the narrower associated CIs, compared with the non-parametric bootstrap method. For the BUN and triglyceride data, the opposite was true. Therefore, when the original

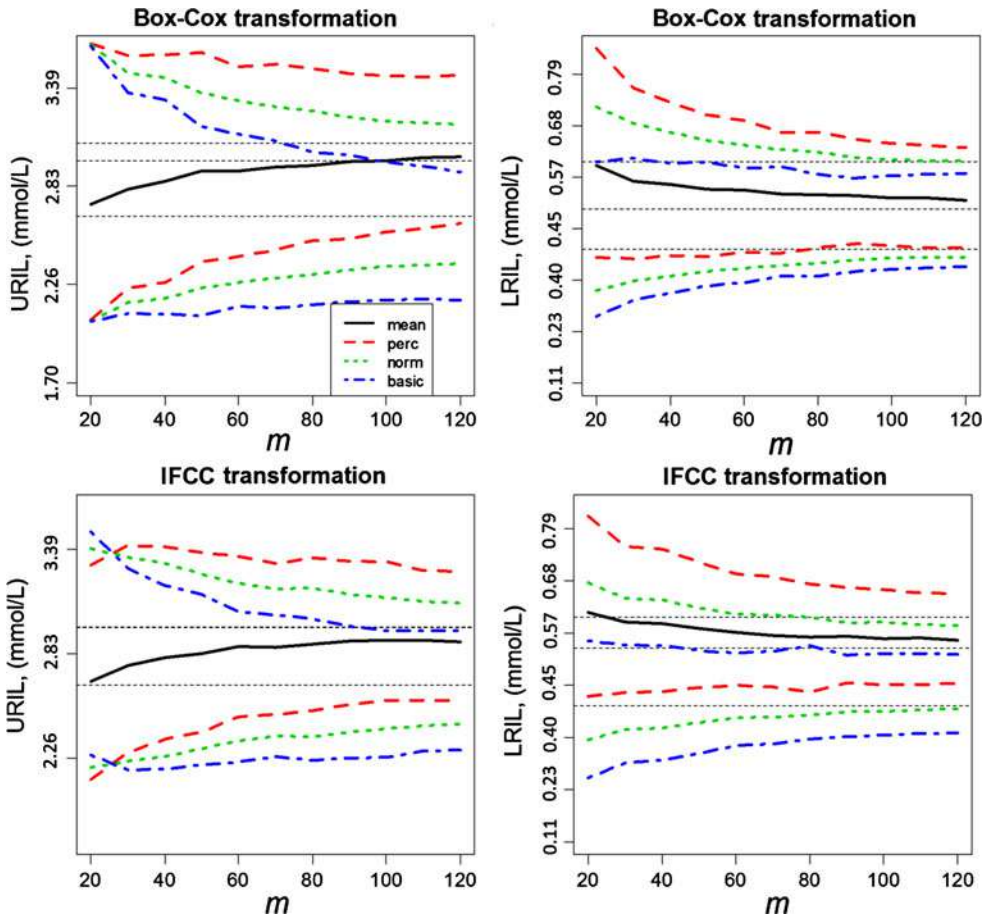

Fig. 5 RI estimates based on the parametric bootstrap method for measurements of triglyceride concentrations with Box-Cox transformation (*top row*) or IFCC recommended transformation (*bottom row*) and with sample sizes $m = 20, 30,\ldots, 120$. For each sample size, 999 bootstrap samples were generated. Horizontal dashed lines are 95 % RIs and the 90 % CIs around the reference limits based on parametric estimation of RIs with 120 observations. Notice that in the upper reference limit with the IFCC transformations, there seem to be two instead of *three dashed lines*; however, it is actually three lines with 2.966 (RI upper limit), 2.653 (lower CI), and 2.972 (upper CI). Legend labeling is as in Fig. 1

data are close to being normally distributed, the parametric bootstrap approach may be preferred over the nonparametric bootstrap method, whereas the nonparametric bootstrap method may be preferable when the original data show a significantly non-normal distribution. This suggests that a transformation to normality does not entirely compensate for the deviation of the original data from normality, but may reduce its impact on the RIs. With bootstrap method (when better of parametric and nonparametric bootstrap is employed), about 60 (in fact 40–60) observations seem to suffice instead of 120 observations required for the nonparametric method with a slight loss in precision, that is, slight increase in the widths of the CIs for the end points of the RIs.

Conclusion

Our findings in this article are suggestive of reduction in the required sample size in RI estimation. Using the methods presented in this study, we can obtain RIs using fewer samples at a cost of slightly more uncertainty around the end points. Nevertheless, this work also points out potential prospective research aspects such as the effect of skewness on nonparametric bootstrap and the level of deviation from normality that would render parametric bootstrap biased. An extensive investigation based on four data sets would not provide very general results, in particular, for nonparametric bootstrap RI estimation. However, based on statistical theory, we can conclude that it is possible to use bootstrap for skewed distributions or even with heavy-tailed distributions, but with attentive care. Nonparametric bootstrap is more robust to skewed distributions, but may require larger samples compared to parametric bootstrap on untransformed data. Our study suggests that even for skewed distributions, nonparametric bootstrap might still require sample sizes smaller than 120. Moreover, with the IFCC transformation procedure, it is possible to transform skewed or heavy-tailed data to approximate normality. Then, one can perform parametric

bootstrap on the transformed data, estimate RIs, and transform back to the original data units. The IFCC transformation usually fixes skewness or heavy-tailedness to a greater extent and then the transformed data are approximately normal. The main problem with transformation method is not the severity of skewness; it is how close it is possible to transform data to normality. The closer the transformed data to normality, the more appropriate the parametric bootstrap methods.

Taken together, we concluded that the nonparametric bootstrap and proposed parametric bootstrap (possibly with transformation to normality) methods are simple, reliable, and particularly cost-effective and can be incorporated into clinical laboratories.

References

1. Henny J (2007) Interpretation of laboratory results: the reference intervals, a necessary evil? Clin Chem Lab Med 45:939–941
2. Solberg HE (2006) In: Burtis CA, Ashwood ER, Bruns DE (eds) Tietz textbook of clinical chemistry and molecular diagnosis, 4th edn. Elsevier Sounders, USA
3. Grasbeck R, Saris NE (1969) Establishment and use of normal values. Scand J Clin Lab Invest 26 Suppl:S110
4. Henny J, Petersen PH (2004) Reference values: from philosophy to a tool for laboratory medicine. Clin Chem Lab Med 42: 686–691
5. CLSI (2008) Defining, establishing, and verifying reference intervals in the clinical laboratory. Approved guideline, 3rd edn. CLSI document C28-A3c, Clinical and Laboratory Standards Institute, Wayne
6. Solberg HE (1993) A guide to IFCC recommendations on reference values. JIFCC 5:160–164
7. International Federation of Clinical Chemistry (IFCC) Panel on Theory of Reference Values, The theory of reference values (1987) Part 5. Statistical treatment of collected reference values. Determination of reference limits. J Clin Chem Clin Biochem 25:645–656
8. Linnet K (1987) Two-stage transformation system for normalization of reference distributions evaluated. Clin Chem 33:381–386
9. Horn PS, Pesce AJ, Copeland BE (1998) A robust approach to reference interval estimation and evaluation. Clin Chem 44: 622–631
10. Linnet K (2000) Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 46:867–869
11. Wright EM, Royston P (1999) Calculating reference intervals for laboratory measurements. Stat Methods Med Res 8:93–112
12. Harrell FE, Davis CE (1982) A new distribution-free quantile estimator. Biometrika 69:635–640
13. Bain LJ, Engelhardt M (2000) Introduction to probability and mathematical statistics, 2nd edn. Duxbury Press, Belmont
14. Elveback LR, Guillier CL, Keating FR Jr (1970) Health, normality, and the ghost of Gauss. J Am Med Assoc 211:69–75
15. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1st edn. Boca Raton
16. Davison AC, Hinkley DV (1997) Bootstrap methods and their application, Cambridge Series in Statistical and Probabilistic Mathematics, 1st edn. Cambridge University Press, New York