Clinical Biochemistry 71 (2019) 14–16

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Choosing the best statistical method for reference interval estimation

Higgins V.[a,b,1], Asgari S.[a,1], Adeli K.[a,b,*]

[a] CALIPER Program, Department of Pediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
[b] Department of Laboratory Medicine & Pathobiology, University of Toronto, Toronto, ON M5S 1A8, Canada

Clinical laboratory tests are often interpreted using reference intervals (RIs; i.e. normative values), which are most commonly defined as the central 95% of values expected in a healthy reference population (1). RI establishment involves defining and sampling a reference population, sample collection, analytical measurement, and statistical analysis. Statistical methods have been described for several aspects of RI establishment, including appropriate outlier detection and removal [2–4], partitioning RIs by covariates (e.g. age and sex) [5,6], and calculating confidence intervals (CIs) around reference limits (1). However, this opinion paper focuses on the direct calculation of RIs from a healthy reference population. The Clinical and Laboratory Standards Institute (CLSI) EP28-A3c guidelines provide recommendations to establish RIs [1]. Nevertheless, a systematic review of RI calculation methods showed that these guidelines are often not followed [7] and we argue that alternative methods from those recommended in the guidelines may be more appropriate in certain situations.

The CLSI EP28-A3c guidelines provide recommendations for calculating RIs using a reference population of healthy individuals, referred to as the direct sampling method [1]. This guideline focuses on two statistical methods: the traditional non-parametric and robust methods. The traditional non-parametric method is recommended when the sample size is ≥120. This method makes no assumptions about the population distribution shape and is based on observation ranks (in order of magnitude), ignoring absolute values. When the sample size is < 120, the robust method of Horn and Pesce [8] is recommended, which is not sensitive to the violation of normality. Although there is no specific minimum observations required, larger sample sizes provide more accurate estimations [1]. The robust method is an iterative process, where the initial center of the data is estimated by the median (i.e. central tendency) and the initial spread is estimated by the median absolute deviation. The sample size cut-off of 120 that distinguishes between using these two methods is, in fact, not based on the accuracy of their estimations, but is the sample size required to calculate 90% CIs non-parametrically [1].

Although the CLSI EP28-A3c guidelines primarily focus on these two methods for calculating RIs, we argue that alternative methods including parametric and Harrell-Davis non-parametric bootstrap methods provide more accurate and precise estimates of the true RI in certain circumstances. The CLSI recommends a pragmatic approach due to the common non-Gaussian distribution of analyte concentration, the simplicity of the recommended non-parametric method, as well as the non-distributional assumptions of both non-parametric and robust methods. Nevertheless, the CLSI EP28-A3c guideline mentions both alternative methods. The parametric method is mentioned alongside discussion of the common requirement to transform data to achieve a Gaussian distribution, involving moderately complex statistical techniques and software [1]. Furthermore, the Harrell-Davis non-parametric bootstrap method is mentioned briefly, in which the recommendation for using the traditional non-parametric method is defended by stating that “the Harrell-Davis method involves more complex computation (and demands more expertise) than the simple nonparametric method recommended in this document”, alongside agreeing with its excellent ability to establish RIs. The guidelines also state the sample size requirement of ≥100 for the Harrell-Davis method and its susceptibility to outliers when the sample size is < 150. The Harrell-Davis non-parametric method is a bootstrap version of the traditional non-parametric method and is substantially more efficient [9].

The parametric method assumes the data have a Gaussian (i.e. normal) distribution and the 2.5th and 97.5th percentiles are determined by the mean +/− 1.96*standard deviation. As analyte concentration is commonly non-Gaussian distributed, prior to relying on a non-parametric method, achieving normality should first be attempted. To determine if a distribution is normally distributed, statistical tests (e.g. Shapiro-Wilk, Anderson-Darling) together with graphical tools (e.g. quantile-quantile (Q-Q) plot) should be used. Normality can often be achieved by reducing skewness and kurtosis using Box-Cox transformation, which determines the optimal value for transformation (i.e. λ) to achieve the best approximation of a normal distribution (e.g. $\lambda = 0$ applies a log transformation). If the Box-Cox transformation fails to achieve normality, the modified Box-Cox transformation can be used [10], which improves the flexibility and performance of the transformation by introducing a corrective term for origin displacement [11]. A recent study systematically compared the performance of the traditional non-parametric, parametric, and robust methods on 216 simulated datasets with various sample sizes and distributions, which were similar to empirical pediatric reference value datasets [12]. According

[*] Corresponding author at: Clinical Biochemistry, DPLM, The Hospital for Sick Children, 555 University Avenue, Toronto, ON M5G 1X8, Canada.
E-mail address: khosrow.adeli@sickkids.ca (K. Adeli).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.clinbiochem.2019.06.006
Received 4 April 2019; Received in revised form 29 May 2019; Accepted 11 June 2019
Available online 12 June 2019

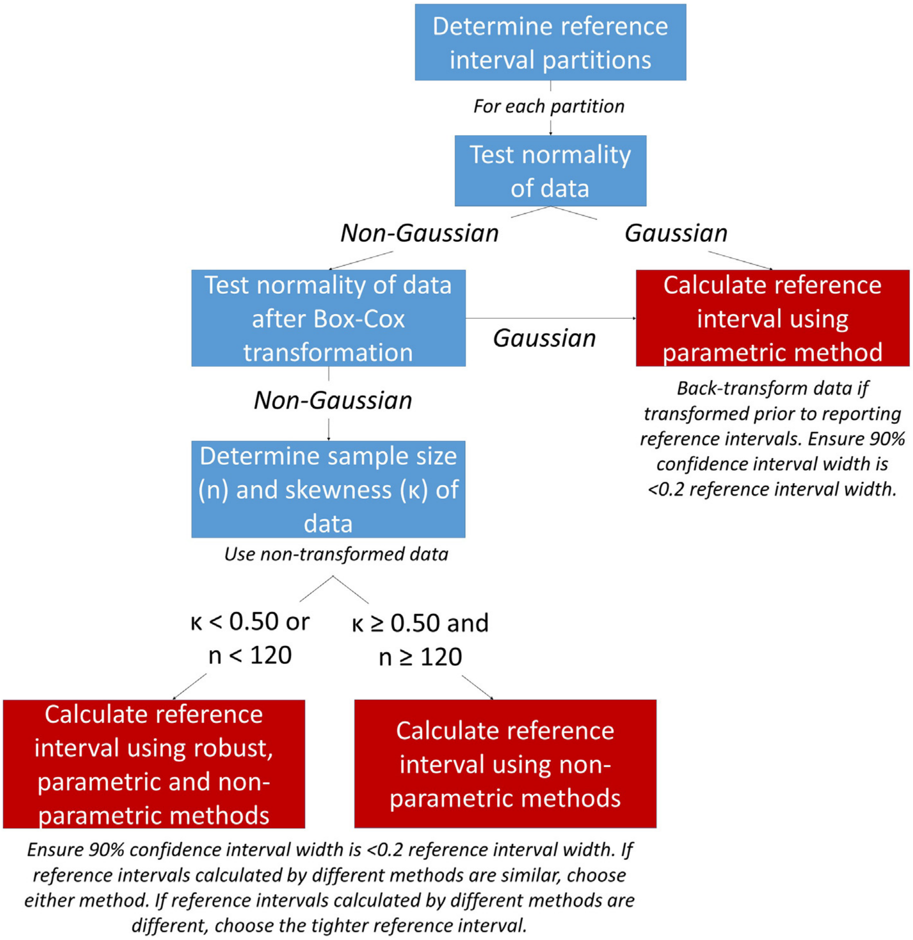
Determine reference interval partitions
For each partition
Test normality of data
Non-Gaussian
Gaussian
Test normality of data after Box-Cox transformation
Gaussian
Calculate reference interval using parametric method
Back-transform data if transformed prior to reporting reference intervals. Ensure 90% confidence interval width is <0.2 reference interval width.
Non-Gaussian
Determine sample size (n) and skewness (κ) of data
Use non-transformed data
κ < 0.50 or n < 120
κ ≥ 0.50 and n ≥ 120
Calculate reference interval using robust, parametric and non-parametric methods
Calculate reference interval using non-parametric methods
Ensure 90% confidence interval width is <0.2 reference interval width. If reference intervals calculated by different methods are similar, choose either method. If reference intervals calculated by different methods are different, choose the tighter reference interval.

Fig. 1. Schematic of the process to select the most accurate and precise method to calculate reference intervals using a reference population of healthy reference individuals.

to this study, the parametric method produces RIs with the least biased (i.e. most accurate) estimates and the lowest mean squared error (i.e. most precise) when the data is Gaussian distributed across all sample sizes tested (i.e. $n = 40–480$) [12]. Accuracy and precision were determined using the pre-defined mean, variance, and skewness of the simulated datasets. For example, using a simulated Gaussian dataset, the upper reference limit empirical bias was lower for the parametric compared to non-parametric method ($n = 40$ −0.039 vs 0.594; $n = 480$: −0.003 vs. 0.048) [12]. Furthermore, using a real Gaussian dataset for uric acid, the parametric method produced narrower CIs than the non-parametric method, particularly at smaller sample sizes (e.g. $n = 120$; parametric: 178 (CI: 159, 196) μmol/L – 455 (437, 472) μmol/L vs. non-parametric: 161 (122, 210) μmol/L – 471 (424, 511) μmol/L) [12].

When data are Gaussian, the precision of RIs calculated with the traditional non-parametric or robust methods increases as sample size increases [12]. However, they are asymptotically biased (unlike those calculated parametrically), meaning the bias does not converge to zero [12], and thus do not produce consistent RI estimates for Gaussian data. Additional studies comparing methods for calculating RIs also recommend the parametric method for Gaussian data or data transformed to Gaussian. Using data from the National Health and Nutrition Examination Survey (NHANES), RIs were calculated using the parametric, Box-Cox transformed parametric, and a non-parametric bootstrapping procedure of resampling with replacement [13]. The transformed parametric method provided the least biased and most precise estimates, followed by the non-parametric bootstrapping method [13]. Furthermore, the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Committee on RIs and Decision Limits (C-RIDL) compared RIs calculated using the parametric method after modified Box-Cox transformation and non-parametric methods and concluded that the parametric method using modified Box-Cox transformation is preferred for RI calculation in most scenarios [14]. Exceptions include when the assay detection limit leads to truncated values on the lower end, in which the non-parametric method may work better for estimating the lower limit [14]. The parametric method may also be advantageous for determining RIs for analytes with commonly observed abnormal values, leading to outlying observations (e.g. triglycerides, alanine aminotransferase) [15]. This is likely due to the greater susceptibility of the non-parametric method to outlying observations, while the parametric method reflects the central portion of the reference distribution (15). However, the parametric method is more sensitive to outliers when the sample size is low and therefore the bootstrap resampling method may be better for these situations [14].

When the reference value distribution is skewed, the best performing method is not as obvious. When the sample size is large ($n \geq 120$), the non-parametric method universally provides the least biased and most precise RI estimate for highly skewed data ($\kappa \geq 0.50$) that cannot achieve normality with Box-Cox transformation [12]. However, the Harrell-Davis non-parametric bootstrap method is recommended for highly skewed data with a large sample size, as this method provides a more accurate and precise estimation of the RI by continuous resampling [16,17]. For example, at a sample size of 500, the root mean squared error (precision estimate) was 8% and 7% lower using the bootstrap compared to traditional non-parametric method when estimating RIs from simulated Gaussian and skewed datasets, respectively [17]. For small values of skewness ($\kappa < 0.50$) and sample size ($n \geq 120$), there is no universal method that provides the most accurate and precise RI estimation [12].

Ideally, RIs should always be calculated using multiple methods, the results compared to each other, and the method that provides the most precise estimate subsequently chosen. One recommended way to determine if the chosen method provides a precise estimation is to ensure the 90% CI is < 0.2 times the width of the 95% RI [16]. Indeed, the width of the CI for each reference limit is a measure of the estimation's precision and represents the sampling variability of the RI. If all methods provide a similar result (reference change value specific for the analyte could be used as a guide to determine a clinically meaningful difference), then the method chosen does not matter. Horn et al. recommended that if the methods provide different results, the RI that is slightly tighter rather than wider, should be chosen to avoid false negative results [18]. Nevertheless, depending on the analyte and its clinical utility, it may not be clear if it is preferable to minimize false negative or false positive results. When finalizing a RI, clinical expertise should be considered to determine appropriate flagging rates for the analyte of interest. Fig. 1 can be used as a guide to help laboratory professionals choose the method most likely to provide the least biased and most precise RI estimate based on characteristics of the data. Given the unique reference value distribution of biochemical markers, the preferred statistical method should be determined for each analyte. It is important for laboratory professionals to understand the statistical methods available to calculate RIs and consequently ensure accurate laboratory test result interpretation.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgments

This work was supported by a Foundation Grant from the Canadian Institutes of Health Research (CIHR) to KA. VH was supported by a CIHR Doctoral award.

References

[1] EP28-A3c, Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory, 3rd edition, Clinical and Laboratory Standards Institute (CLSI), 2010.
[2] A.H. Reed, R.J. Henry, W.B. Mason, Influence of statistical method used on the resulting estimate of normal range, Clin. Chem. 17 (4) (1971 Apr) 275–284.
[3] W. Dixon, Processing date for outliers, Biometrics. (9) (1953) 74–89.
[4] J. Tukey, Exploratory Data Analysis, Addison-Wesley, Boston, 1977.
[5] A. Lahti, P.H. Petersen, J.C. Boyd, P. Rustad, P. Laake, H.E. Solberg, Partitioning of nongaussian-distributed biochemical reference data into subgroups, Clin. Chem. 50 (5) (2004 May) 891–900.
[6] E.K. Harris, J.C. Boyd, On dividing reference data into subgroups to produce separate reference ranges, Clin. Chem. 36 (2) (1990 Feb) 265–270.
[7] C.H. Daly, X. Liu, V.L. Grey, J.S. Hamid, A systematic review of statistical methods used in constructing pediatric reference intervals, Clin. Biochem. 46 (13–14) (2013 Sep) 1220–1227.
[8] P. Horn, A. Pesce, Reference Intervals: A user's Guide, AACC Press, Washington, DC, 2005.
[9] F.E. Harrell, C.E. Davis, A new distribution-free quantile estimator, Biometrika. 69 (3) (1982) 635–640.
[10] K. Ichihara, T. Kawai, 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. 10 (2) (1996) 110–117.
[11] K. Ichihara, J.C. Boyd, IFCC committee on reference intervals and decision limits (C-RIDL). An appraisal of statistical procedures used in derivation of reference intervals, Clin. Chem. Lab. Med. 48 (11) (2010 Nov) 1537–1551.
[12] C.H. Daly, V. Higgins, K. Adeli, V.L. Grey, J.S. Hamid, Reference interval estimation: methodological comparison using extensive simulations and empirical data, Clin. Biochem. 50 (18) (2017 Dec) 1145–1158.
[13] I.Y. Pavlov, A.R. Wilson, J.C. Delgado, Reference interval computation: which method (not) to choose? Clin Chim Acta Int J Clin Chem. 413 (13–14) (2012) 1107–1114 Jul 11.
[14] K. Ichihara, Y. Ozarda, J.H. Barth, G. Klee, L. Qiu, R. Erasmus, et al., A global multicenter study on reference values: 1. Assessment of methods for derivation and comparison of reference intervals, Clin Chim Acta Int J Clin Chem. 467 (2017 Apr) 70–82.
[15] K. Ichihara, Statistical considerations for harmonization of the global multicenter study on reference values, Clin Chim Acta Int J Clin Chem. 432 (2014) 108–118 May 15.
[16] E.K. Harris, J.C. Boyd, Statistical Bases of Reference Values in Laboratory Medicine, Marcel Dekker, New York, 1995, pp. 1–61.
[17] K. Linnet, Nonparametric estimation of reference intervals by simple and bootstrap-based procedures, Clin. Chem. 46 (6) (2000 Jun) 867–869 Pt 1.
[18] P.S. Horn, A.J. Pesce, B.E. Copeland, A robust approach to reference interval estimation and evaluation, Clin. Chem. 44 (3) (1998 Mar) 622–631.