Comparison of Methods to Generate Reference Limits

Bipasa Biswas[1], Nairita Goshal[2]
[1] CDRH, FDA, 10903 New Hampshire Avenue, Silver Spring, MD 20993
[2] University of Illinois at Chicago, 1603 W Taylor Street, Chicago, IL 60612

Abstract
Clinical laboratory tests often require a reference interval for quantitative tests and the construction of such intervals are common for laboratory tests. Reference Intervals are used to determine unusual or extreme measurements in laboratory medicine. The reference interval is the central interval bounded by the reference limits. *In vivo* diagnostic devices in ophthalmology and neurology often compare an individual patient's medical device output/test result against a database of output/test results from subjects deemed to be in good health, for clinical management of the individual patient. The device output/test results from the subjects in good health constitute a reference database (also commonly known as normative database). Reference database for *In Vivo* diagnostic devices is composed of measurements of multiple anatomical or physiological features from healthy individuals. 1st, 2nd, 2.5th, 5th, 95th, 97.5th, 98th or 99th percentiles are usually reported from the reference database. This presentation compares three common methods – Nonparametric, Harrell-Davis and the Robust method to generate reference limits.

Key Words: Reference intervals, reference limits, non-parametric.

1. Reference Database

In laboratory medicine, reference intervals are used commonly to determine unusual or extreme measurements. The reference interval is defined by the interval between two percentile values, centered about the median on the probability scale. The guidance document CLSI EP28-A3c[1] written for Clinical Chemistry describes how to perform reference interval study in detail. The guidance provides the definition of apparently healthy population, statistical methods to calculate reference interval/cut-off based on percentiles, reasons for partitioning based on covariates and describes transfer of reference interval in case a valid reference study already exits.

Reference databases for *In Vivo* diagnostic devices consists of measurements of one or more parameters of an anatomical or physiological feature from reference individuals. Reference databases are a sample of reference individuals usually consisting of cross-sectional (single time point) measurements of either one or more anatomical or

[1]CLSI. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

physiological feature on these individuals. A reference individual is a subject/person selected for testing based on well-defined criteria like the person's state of health and usually individuals are in good health. A reference value is the value or measurement obtained by measurement of a particular anatomical or physiological feature on a reference individual and the distribution of these values constitute a reference distribution.

1.1 Reference limits

The reference limits are often generated from a cross-sectional (i.e. one measurement per parameter in an individual) reference database. A cross-sectional reference database provides information about the variability across multiple individuals at a single time point for the anatomical/physiological feature(s) of interest. Reference limits are values derived from the reference distribution and used for descriptive purpose.

Percentiles (1^{st}, 5^{th}, 95^{th} or 99^{th}) are usually reported from these reference databases. Often, 2.5^{th} and 97.5^{th} percentiles, defining a 95% reference interval (reference interval sometimes also called reference range) are reported from these reference databases or the database is used to generate a z-score which is a commonly seen for Neurological medical devices.

1.2 Examples

Reference databases are common to Optical Coherence Tomography (OCT) devices and neurological medical devices. OCT is a type of imaging device that uses light to capture micrometer-resolution, three-dimensional images from within optical scattering media (e.g., biological tissue). Optical coherence tomography is based on low-coherence interferometry, typically employing near-infrared light. OCT is used in ophthalmology and optometry to obtain detailed images from within the retina which offers objective and quantitative anatomical measurements of the eye. Measurements involve- the topography of the optic nerve head, including the rim and the cup, the peripapillary RNFL, retinal ganglion cell thickness, macular thickness –often in multiple sectors of the eye (superior, inferior, nasal, temporal)

Reference databases in neurological medical devices are often for cognitive battery tests and quantitative EEG based tests. The cognitive battery tests are based on response to sequence of questionnaires where the measurements obtained are on speed of processing, attention/vigilance, working memory, verbal learning, visual learning, reasoning and problem solving, social cognition and overall composite score.
The quantitative EEG is used to record brain's spontaneous electrical activity over a period of time which involves placement of electrodes on the scalp and measurements are non-invasive. The measurements obtained are EEG spectra, behavioral data (omission, commission errors, reaction time and variance of response in the task), and ERP (Evoked Response Potential) independent components.

2. Statistical Methods to generate reference limits

2.1 Statistical methods

The document CLSI EP28-A3c, written for Clinical Chemistry discusses statistical methods for estimating reference interval and percentiles. The different methods for estimating reference limits discussed are -Nonparametric methods, Iterative weighted percentile method (also known as robust method), Harrell-Davis method and Bootstrap Method. In this presentation three methods- Nonparametric, Harrell-Davis and the Robust

methods are studied and compared. Data sets following specific distributions were simulated to observe the performance of the three methods.

2.1.1 Non-parametric

Nonparametric method is a distribution-free method using rank of ordered observed values. The ordered observations are denoted as $x_{(1)}, x_{(2)}, \ldots, x_{(r)}, \ldots; x_{(n)}$. The p^{th} sample quantile Q_p is the observation with rank $r = p(n+1)$. In case of non-integer r values, quantiles are obtained by linear interpolation.

2.1.2 Harrell-Davis Method

Harrell-Davis method (6) is based on a linear combination of the order statistics using difference between two incomplete beta functions as weight. The p^{th} percent sample quantile is expressed as

$$Q_p = \sum_{i=1}^{n} W_{n,i} X_{(i)}$$

Where

$$W_{n,i} = \frac{1}{\beta\big((n+1)p, (n+1)(1-p)\big)} \int_{(i-1)/n}^{i/n} y^{(n+1)p}(1-y)^{(n+1)(1-p)^{-1}} dy$$
$$= I_{i/n}[p(n+1), (1-p)(n+1)] - I_{(i-1)/n}[p(n+1), (1-p)(n+1)]$$

Here $I_x(a,b)$denotes the incomplete beta function.

2.1.3 Robust Method

This method uses robust estimates (8) of location and scale with iterative bi-weight approach. The $(1-\alpha/2)100$ % bi-weight reference interval for a symmetric distribution is

$$T_{bi}(C_1) \pm t_{n-1}(1-\alpha/2)\sqrt{S_T^2(C_1) + S_{bi}^2(C_2)}$$

Where $T_{bi}(C_1)$ is the bi-weight location estimator with tuning constant C_1, $t_{n-1}(1-\alpha/2)$ is the quantile from Students t-distribution with n-1 degrees of freedom, $S_T^2(C_1)$ is the bi-weight estimator of the variability of T_{bi} and $S_{bi}^2(C_2)$ is the bi-weight estimator of the spread with tuning constant C_2. For observations $x_1, x_2, \ldots, x_n$ the bi-weight estimator of location T_{bi} is defined as the solution to the equation

$$\sum_{i=1}^{n} \Psi(u_i) = 0$$

where

$$u_i = (x_i - T_{bi})/cs_{bi}$$
$$\Psi(u_i) = u_i\, w(u_i)$$
$$w(u) = (1-u^2)^2;\ |u| < 1$$
$$= 0\ otherwise$$

s_{bi} is a bi-weight estimate of spread and c is a tuning constant. Solving the above yields

$$T_{bi} = \frac{\sum w(u_i) x_i}{\sum w(u_i)}$$

T_{bi} is computed by iteration, using the above formula, with median as initial value and initial estimate of spread is median absolute deviation (MAD) divided by 0.6745. T_{bi} is updated until change in consecutive iterative values are negligible (<0.001).

3. Simulations

To compare the three methods, a simulation using different distributions type – Gaussian, right skewed, left skewed and a heavy tailed symmetric distribution were performed. A random sample of size 240 were generated from each of the distributions - Symmetric - Normal(9.63,1.2) Distribution, Right Skewed-Gamma(1,2) distribution, Left Skewed-Beta(5,1) Distribution, Heavy Tailed- t distribution with 2 degrees of freedom.

A sample size of 240 was fixed for estimating the percentiles -2.5th, 5th, 95th, and 97.5th.by the three methods (Nonparametric, Harrell-Davis, and Robust). Each type of distribution mentioned above were simulated 1000 times with a sample size of 240 and from each simulation, estimates of 2.5th, 5th, 95th, and 97.5th were generated. The bias of these estimates was checked against the actual values from these distributions in a box plot. Figures 1 through 4 are the side by side box plots of the three methods for the distributions of 2.5th, 5th, 95th, and 97.5th percentiles for the 1000 simulations of sample size 240 from each of the four distribution types.

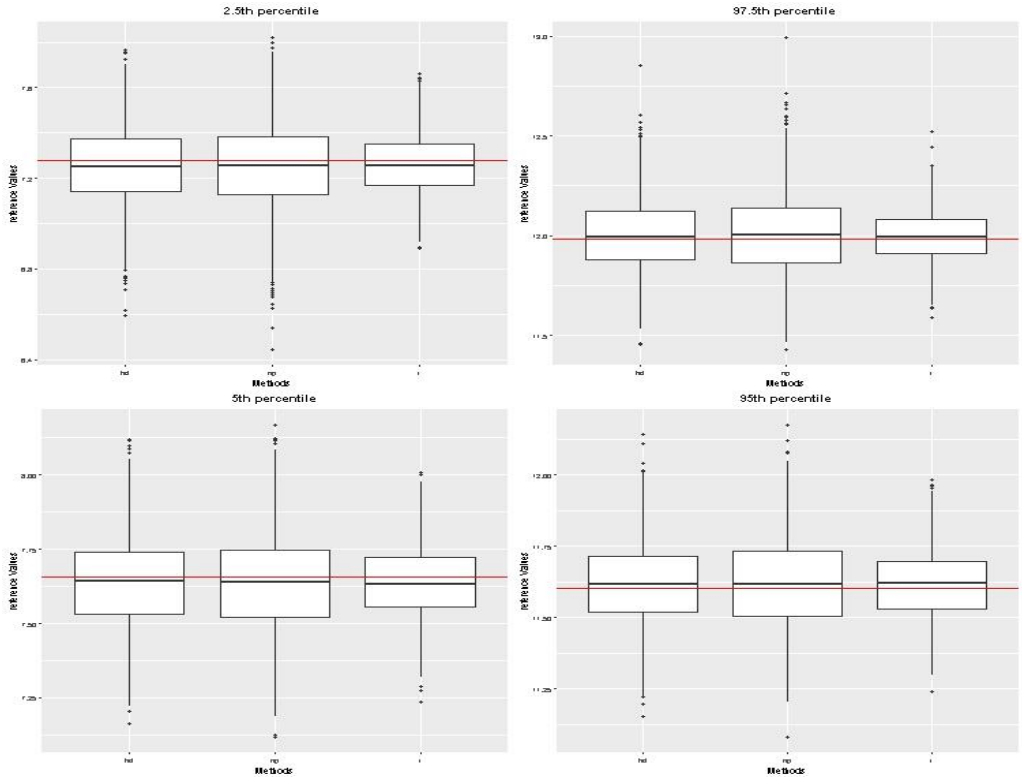

Figure 1: Estimated percentiles from Gaussian (N(9.63,1.2)) distribution by the three methods with the true percentile represented by the red horizontal line.

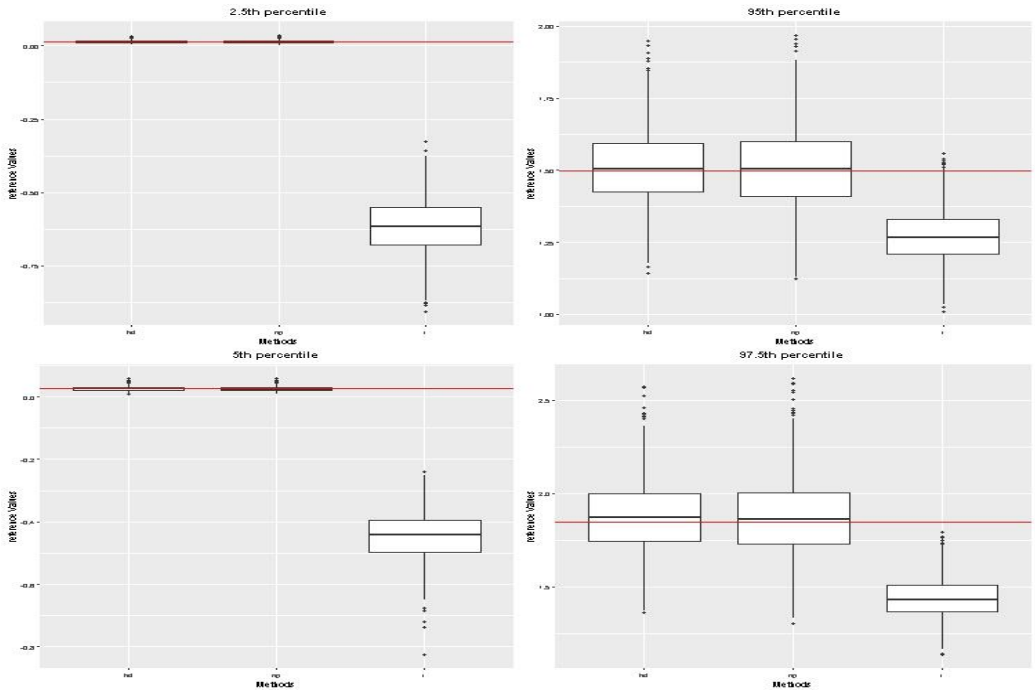

Figure 2: Estimated percentiles from Gamma (1,2) distribution by the three methods with the true percentile represented by the red horizontal line.

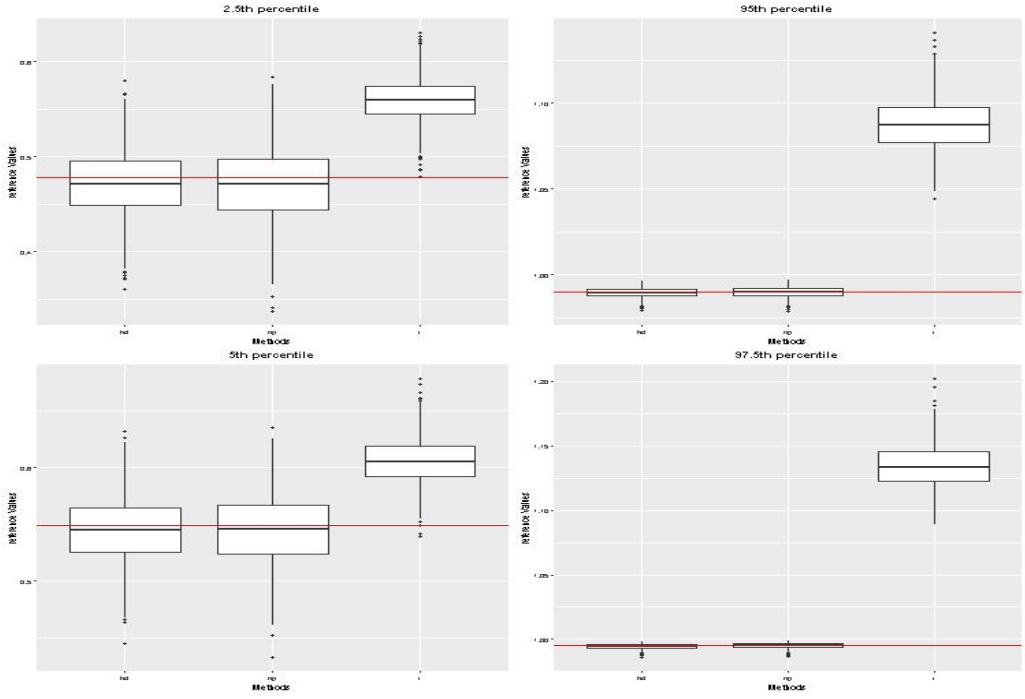

Figure 3: Estimated percentiles from Beta(5,1) distribution by three methods with the true percentile represented by the red horizontal line.

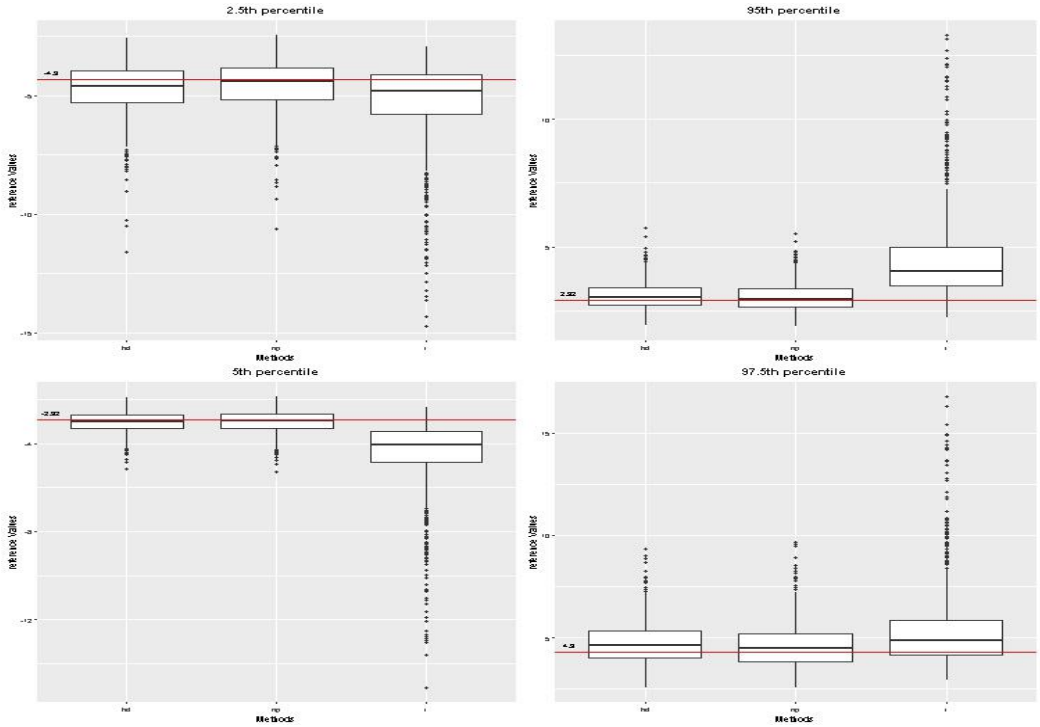

Figure 4: Estimated percentiles from t(2) distribution by the three methods with the true percentile represented by the red horizontal line.

Further to compare the estimated percentiles (2.5^{th}, 5^{th}, 95^{th}, 97.5^{th}) from the three methods, for four different distributions of reference values, the root mean square error of each of the estimated percentiles were compared for varying sample sizes. 1000 simulations for each of the four distributions types Symmetric - Normal(9.63,1.2) Distribution, Right Skewed-Gamma(1,2) distribution, Left Skewed- Beta(5,1) Distribution, Heavy Tailed- t distribution with 2 degrees of freedom , were generated for varying sample sizes from 50,80,120, 240, and 720, and the four percentiles 2.5^{th}, 5^{th}, 95^{th}, 97.5^{th} by each method were generated. The root mean square error (RMSE) was calculated using

$$RMSE = \sqrt{Bias^2 + STD^2}$$

Where the bias is the difference of the mean of the estimated percentiles the 1000 simulations for each distribution type minus the true value of the percentile for that distribution and the standard deviation is the standard deviation of the 1000 estimated percentiles for each distribution. The figures 5 through 8 plot the RMSE against the sample size for the three methods by each of the four distribution type.

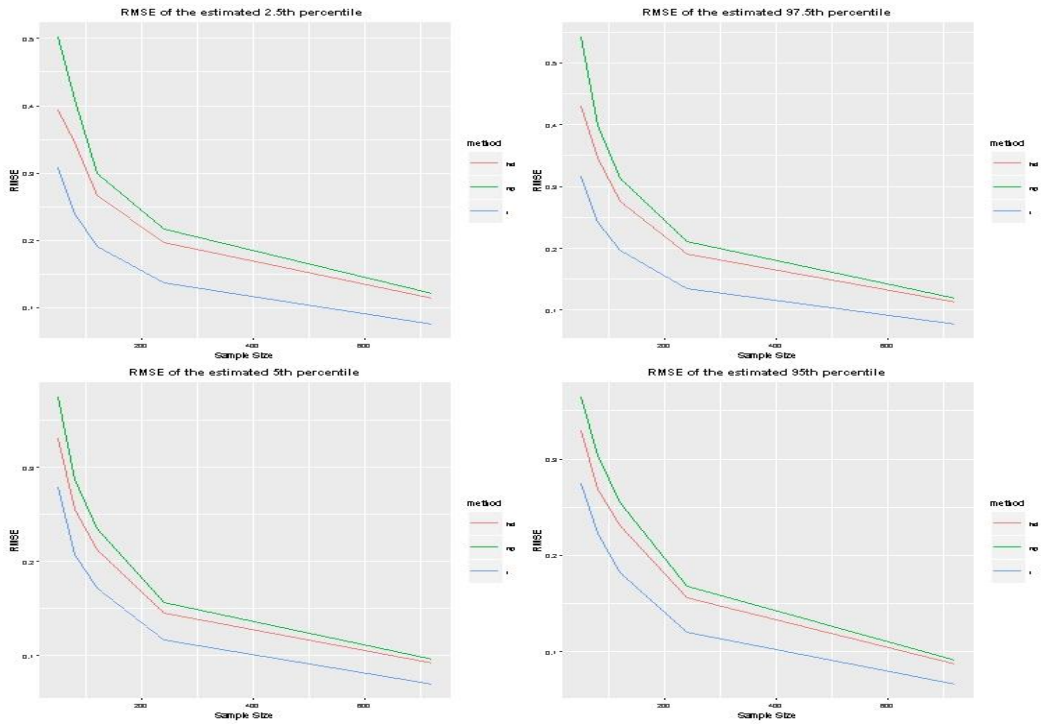

Figure 5: RMSE for estimated percentiles by sample size (N(9.63, 1.2))

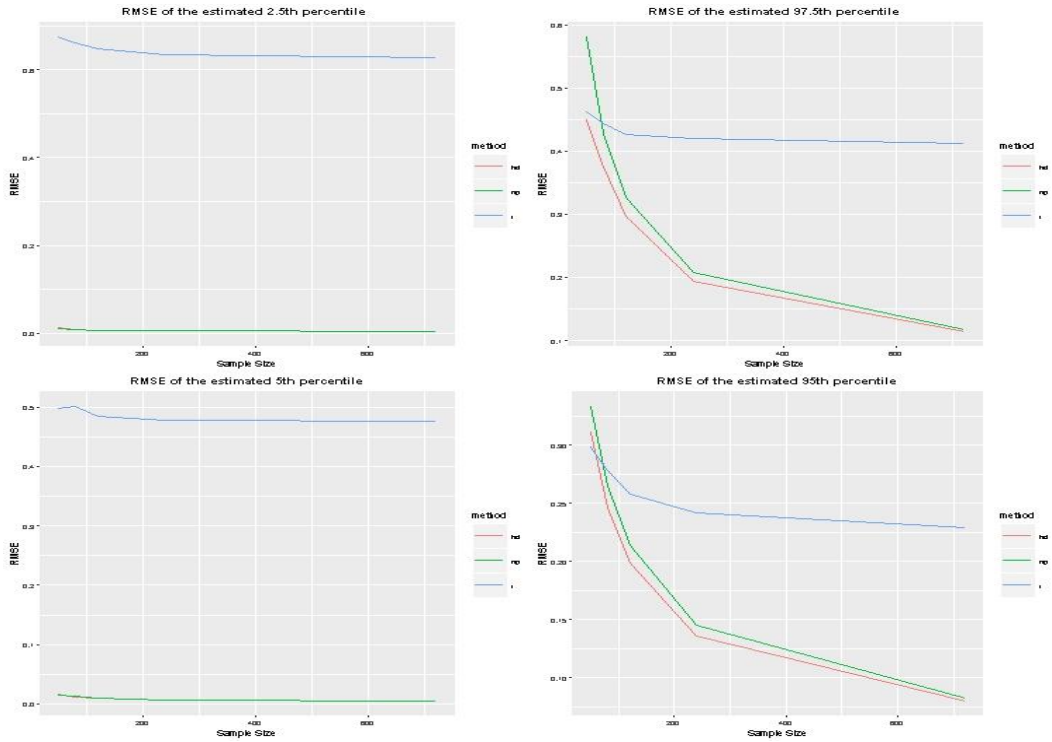

Figure 6: RMSE for estimated percentiles by sample size (Gamma(1,2))

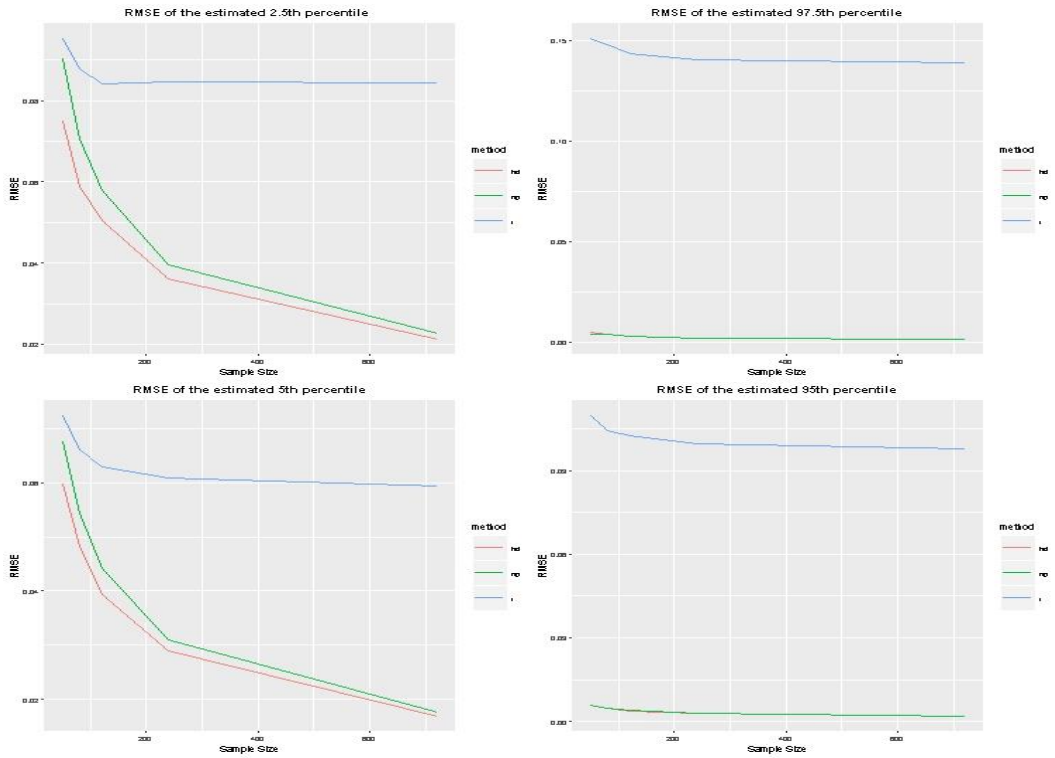

Figure 7: RMSE for estimated percentiles by sample size (Beta(5,1))

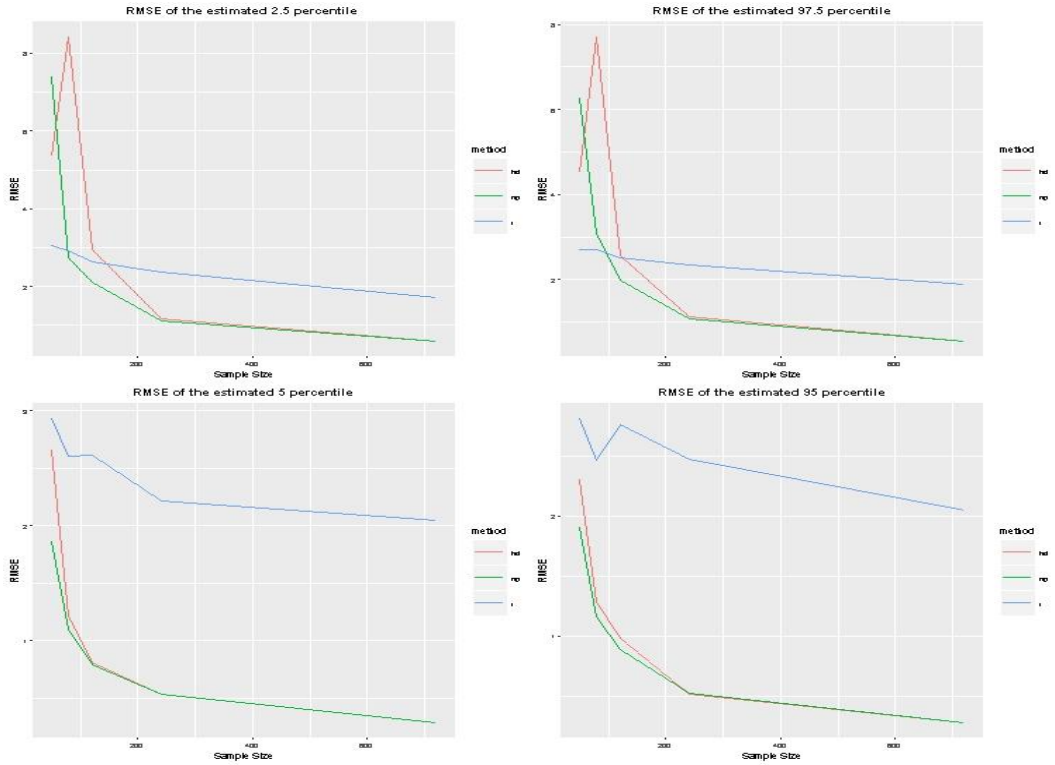

Figure 8: RMSE for estimated percentiles by sample size (t(2))

4. Coverage Probability

Coverage probability of 95% confidence intervals at 2.5th and 97.5th percentiles by the three methods fixing sample size at 146 were further evaluated to assess the performance of the confidence intervals for the three methods for each of the four distribution types evaluated in the presentation. The sample size of 146 was selected based on width of the 95% confidence interval of nonparametric estimate of 2.5th percentile.

Table 1: Coverage probability of 95% confidence interval for the percentiles generated by the three methods.

	Nonparametric		Harrell-Davis		Robust	
	2.5th	97.5th	2.5th	97.5th	2.5th	97.5th
Normal Distribution	0.94	0.93	0.90	0.90	0.94	0.93
t Distribution	0.95	0.94	0.91	0.91	0.81	0.82
Gamma Distribution	0.95	0.94	0.91	0.89	0.00	0.25
Beta Distribution	0.93	0.93	0.90	0.91	0.37	0.00

5. Conclusions

Reference database for *In Vivo* diagnostic devices often involve multiple parameter measurements per individual and these databases constitutes of measurements from reference individuals, also referred to as reference values.

Three statistical methods- nonparametric, Harrell-Davis and the Robust method, were compared with different distributions of reference values. In particular four distributions types - Symmetric - Normal(9.63,1.2) Distribution, Right Skewed-Gamma(1,2) distribution, Left Skewed- Beta(5,1) Distribution, Heavy Tailed- t distribution with 2 degrees of freedom, were used to compare the effect on the estimates of the reference limits by the three methods by evaluating bias, root mean square error by varying sample size and the coverage probability of the 95% confidence interval of the estimated reference limits.

Based on the simulations, the box plots (Figure 1) show robust method is more efficient if the underlying distribution of reference values is normal (or Gaussian) and provides a savings in sample size (Figure 5). However, if the distribution deviates from normality, the non-parametric method and Harrell-Davis provides an unbiased estimate of the reference limits (Figures 2-4). Evaluations based on RMSE indicate that Robust method results in a savings in sample size if the underlying distribution of reference values is normal. However, for non-normal distribution, the nonparametric method provides better overall estimate of the reference limits based on the evaluation of RMSE (Figures 6-8).

The coverage probability of the 95% confidence intervals of the reference limits by the three methods also show better coverage for non-parametric method.

References

1. Realini, T, Zangwill, L, Flanagan, J, Garway-Heath, D, Patella, V M, Johnson, C, Artes, P, Gaddle I B, Fingeret, M. (2015) Normative Databases for Imaging Instrumentation. *J Glaucoma* 24(6): 480-483. doi: 10.1097/IJG.0000000000000152.
2. Kern, RS, Nuechterlain, KH, Green, MF, Baade, LE, Fenton, WS, Gold, JM, Keefe, RSE, Mesholam-Gately, R, Mintz, J, Seidman, LJ, Stover, E, Marder, SR. (2008) The MATRICS Consensus Cognitive Battery, Part 2 Co-Norming and Standardization. *Am J Psychiatry* 165:214-220.
3. CLSI. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.
4. Linnet, K. (2000) Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. *Clinical Chemistry* 31: 867-869.
5. Crawford, JR, Howell, DC. (1998) Comparing an Individual's Test Score against norms Derived from Small Samples. *J Clinical Neuropsychologist* 12(4): 482-486.
6. Harrell, FE, Davis, CE. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3):635-640.
7. Hahn, G., & Meeker, W. (1991). Statistical intervals: a guide for practitioners. New York: John Wiley & Sons.
8. Horn, PS. (1988) A bi-weight prediction interval for random samples. *J Am Stat Assoc.* 83:249-256.
9. Horn, P, Pesce, A, Copeland, B. (1998). A robust approach to reference interval estimation and evaluation. *Clinical Chemistry*, 44(3): 622-631.
10. Horn, P, Pesce, A, & Copeland, B. (1999). Reference interval computation using robust vs. parametric and nonparametric analyses. *Clinical Chemistry*, 45(12):2284-2285.
11. Beal D. SESUG (2012). Sample size determination for nonparametric upper tolerance for any order statistic.
12. Biswas, B. ENAR (2017) Reference Databases for In Vivo Diagnostic Devices.