Reference IntervalsComparison of Calculation Methods and Evaluation of Procedures for Merging Reference Measurements From Two US Medical Centers

George G Klee, MD, PhD, Kiyoshi Ichihara, MD, PhD, Yesim Ozarda, MD, Nikola A Baumann, PhD, Joely Straseski, PhD, Sandra C Bryant, MS, Christina M Wood-Wentz, MS

Reference Intervals

Comparison of Calculation Methods and Evaluation of Procedures for Merging Reference Measurements From Two US Medical Centers

George G. Klee, MD, PhD,[1] *Kiyoshi Ichihara, MD, PhD,*[3] *Yesim Ozarda, MD,*[4] *Nikola A. Baumann, PhD,*[1] *Joely Straseski, PhD,*[5] *Sandra C. Bryant, MS,*[2] *and Christina M. Wood-Wentz, MS*[2]

From the [1]Department of Laboratory Medicine and Pathology and [2]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN; [3]Yamaguchi University Graduate School of Medicine, Ube, Japan; [4]Uludag University, Bursa, Turkey; and [5]University of Utah, Department of Pathology, Salt Lake City.

Key Words: Reference values; Normal values; Merging reference data; Method comparison; Serum panel; Parametric method; Nonparametric method; Box-Cox power transformation; Latent abnormal values exclusion (LAVE)

Am J Clin Pathol December 2018;150:545-554

DOI: 10.1093/AJCP/AQY082

ABSTRACT

Objectives: *To analyze consistency of reference limits and widths of reference intervals (RIs) calculated by six procedures and evaluate a protocol for merging intrainstitutional reference data.*

Methods: *The differences between reference limits were compared with "optimal" bias goals. Also, widths of the RIs were compared. RIs were calculated using Mayo-SAS quantile, EP Evaluator, and four International Federation of Clinical Chemistry and Laboratory Medicine methods: parametric and nonparametric (NP) with and without latent abnormal values exclusion (LAVE). Regression parameters from cotested samples were evaluated for harmonizing intrainstitutional reference data.*

Results: *Mayo-SAS quintile, LAVE(−)NP, and EP Evaluator generated similar RIs, but these RIs often were wider than RIs from parametric procedures. LAVE procedures generated narrower RIs for nutritional and inflammatory markers. Transformation with regression parameters did not ensure homogeneity of merged data.*

Conclusions: *Parametric methods are recommended when inappropriate values cannot be excluded. The nonparametric procedures may generate wider RIs. Data sets larger than 200 are recommended for robust estimates. Caution should be exercised when merging intrainstitutional data.*

All laboratories should have valid reference intervals (RIs) for their assays to help clinicians interpret test results. Two questions often arise when considering the establishment of the intervals: (1) What statistical methods or programs should be used to calculate these intervals? and (2) Is there a way to use reference data from another institution? The Clinical and Laboratory Standards Institute (CLSI) has developed a 57-page guideline (CLSI EP28-193C:2010) for *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory*.[1] This is a comprehensive document but may be too complex for many laboratories. This guideline recommends using a minimum data set of 120 healthy participants for establishing RIs for each important subgroup, such as sex and race. They discuss the topic of multicenter RI studies but recommend that each laboratory should determine its own RIs because of analytic differences and population differences. They do not discuss methods for transferring reference data between laboratories.

The issue of local population differences may be important in some specialized practices, but its relevance for general medical centers in the current age of patient mobility and population heterogeneity is debatable. Most medical centers see both local patients and patients from other localities. In the United States, many communities comprise individuals from various ethnic backgrounds. Even individuals with similar ancestral roots may not

comprise a homogeneous population of laboratory values due to dietary and environmental variables.[2]

In this article, we compare six methods for generating RIs. Two of the methods are commercially available, and the other four methods were developed under the guidance of the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Committee on Reference Intervals and Decision Limits (C-RIDL). We also evaluate an algorithm to transform reference value data collected at one site for use at a second medical center based on using a panel of reference samples to compensate for the analytic differences between centers.[3]

Materials and Methods

Volunteer healthy donors were recruited at two medical centers (ARUP Laboratories, Salt Lake City, UT, and Mayo Clinic, Rochester, MN) using a standardized study design and a locally adapted protocol approved by the institutional review board. The inclusion criteria and blood collection protocol were endorsed by the IFCC C-RIDL.[4] Inclusion criteria were that participants felt subjectively well and were 18 years or older. The following medications were permitted, but the name, dose, and frequency were recorded: contraceptive pills, estrogens, and thyroxine. If the donor was taking thyroxine, his or her thyroid-stimulating hormone (TSH) level was to be less than the laboratory's current upper reference limit. Exclusion criteria were known diabetics, history of chronic liver or kidney disease, hospitalized or serious illness within the previous 4 weeks, blood donation in the previous 3 months, pregnant or within 1 year of childbirth, or other significant disease. The target sample size was 240, with about half in each sex. In this study, 92 men and 148 women were collected at Mayo, whereas 125 men and 125 women were collected at ARUP. Participants were requested to avoid excessive physical activity for 3 days before blood collection. After obtaining informed consent, blood was collected and processed at both sites according to the C-RIDL protocol.[4]

The RI analytes were measured shortly after collection at the ARUP site, whereas aliquots were frozen at –80°C and analytes were measured on 3 consecutive days at the end of the sample collection at Mayo. Roche Modular (Indianapolis, IN) analyzers were used to measure the general chemistry analytes at both sites. Carcinoembryonic antigen (CEA), cortisol, and prolactin (PRL) were measured on the Siemens Centaur (Walpole, MA) at ARUP, whereas CEA was measured on the Beckman Coulter DxI (Chaska, MN), and cortisol and PRL were measured on Roche Modular at Mayo. Parathyroid hormone (PTH) and TSH were measured on the Roche E170 at ARUP, whereas they were measured on the Roche Modular at Mayo.

Procedures for Calculation of RIs

The Mayo-SAS quantile program uses the SAS QUANTREG quantile regression procedure to calculate specific percentiles and analyze the effects of secondary variables such as age and sex.[5] Ordinary least squares regression models optimize the best fit for the conditional mean of the response variable for a given conditional variable. Quantile regression extends this regression model to optimize the best fit of conditional quantiles (such as the 2.5th and 97.5th percentiles) of the response variable.[6] The main advantage of quantile regression is for modeling data with heterogeneous conditional distributions since it makes no distributional assumption about the error term[7] and thus belongs to the nonparametric method in a broad sense. One sample from a 97-year-old man was excluded as an outlier from the Mayo-SAS quantile analysis. The standard errors of the quantile estimates were calculated using a bootstrap resampling procedure with replacement with 10,000 replicates.

EP Evaluator is a widely used commercial Microsoft Windows–based software program marketed by Data Innovations (South Burlington, VT) to facilitate statistical analyses in clinical laboratories.[8] It incorporates many of the recommendations of the CLSI. It has procedures both to establish and to verify RI. This system can partition data sets by any of the included variables such as sex, but you can partition by only one variable at a time. Differences by age can be analyzed by using multiple age bins if the sample size is large enough. Bounds can be set to exclude extreme values. The program offers nonparametric, parametric, and transformed parametric statistics. The system uses a Box-Cox transformation based on an exponent and a constant. It tests to see if transformations significantly improve the fit to the Gaussian distribution. We partitioned the data by sex and produced separate results for women and men. Since all the female subgroups had at least 120 values, the system made nonparametric estimates for each analyte for women. However, the men from Mayo had fewer than 120 values, and the system used transformed parametric statistics to estimate the RIs for all but six analytes (albumin, calcium, chloride, potassium, magnesium, sodium), which were based on Gaussian parametric estimates.

The four IFCC procedures for calculating RIs are comprehensive methods for parametric and nonparametric estimates with and without exclusion of inappropriate values. The exclusion criterion is based on the latent abnormal values exclusion (LAVE) method.[1,9] This method estimates values from participants with abnormal values in related analytes caused by common disorders such as metabolic syndrome, muscular damage, and inflammation. The reference analytes actually used in the procedure for identifying inappropriate values were uric acid, glucose, triglycerides, aspartate aminotransferase (AST), alanine aminotransferase (ALT), lactate

DOI: 10.1093/ajcp/aqy082

dehydrogenase (LDH), γ-glutamyltransferase (GGT), and creatine kinase (CK), which were sensitive to the above conditions. The LAVE procedure is superior to outlier exclusion because it does not truncate the reference distribution. The disadvantage of using the exclusion criteria is the reduction in the number of values that can be used to define the RIs. The parametric method is based on the two-parameter (modified) Box-Cox equation, which includes a parameter representing an origin of transformation for improved fitting to the Gaussian shape.[10]

$$X = \frac{(x-a)^p - 1}{p} \cdots p \neq 0.0$$

$$X = \log(x-a) \cdots p = 0.0$$

where x and X denote observed value before and after transformation, and the parameters p and a, respectively, represent power and the origin of transformation to be set below a minimal value of x. The classic Box-Cox method does not have the parameter a. The parametric method features truncation of values outside mean ± 2.57 SD at the transformed scale (1% on both ends of the distribution), which is effective in identifying extreme values unmatched to the central shape of the distribution. In all the IFCC procedures, the 90% confidence intervals for the reference limits were generated by the bootstrap method with 100 repetitive resampling.

Procedures for Comparing Reference Limits and Interval Widths

Separate analyses were performed for women and men to compare the six methods for calculating RIs and confidence intervals for each analyte. The RI limits for both the lower limits (LLs) and upper limits (ULs) for both men and women were separately derived using each statistical protocol. The differences between these limits derived from each of the six statistical protocols were compared pairwise, resulting in 15 combinations. A procedure proposed by Ozarda et al[11] was used to normalize the reference limit differences. This procedure compares the absolute difference between the reference limit estimates divided by between-individual standard deviations roughly calculated from the reference ranges:

$$\text{LL Ratio} = \left| \text{LL}_\text{Y} \quad \text{LL}_\text{Y} \right| / \left(\text{UL}_\text{Y} \quad \text{LL}_\text{Y} \right) / 3.92$$

$$\text{UL Ratio} = \left| \text{UL}_\text{X} \quad \text{UL}_\text{Y} \right| / \left(\text{UL}_\text{Y} \quad \text{LL}_\text{Y} \right) / 3.92$$

where X is the calculation method being evaluated and Y is the comparator method.

By analogy to the theory of acceptable analytic bias in laboratory tests, the "optimal limits" for analytic bias are set at 12.5% of combined individual and group biologic coefficient of variation.[12,13] When the optimal values were not available, we used half of the desirable bias limits as equivalent to the optimal limits.[13] The upper and lower reference limits were calculated for both male and female participants using the six methods. The reference limits derived from each pair of methods were compared pairwise. The ratio of the absolute differences between the index method reference limit vs the alternate reference limit was compared with the absolute reference range derived from the index method. These ratios were converted to percentages. The methods were considered divergent when the ratio percentage exceeded the "optimal limits" for analytic bias.

The effects of extreme values on the widths of the RIs were evaluated for each of the statistical calculation procedures. A critical value equivalent to UL ratio (or LL ratio) was computed as a ratio of absolute differences in average UL (or LL) between parametric (P) and nonparametric (NP) methods to the average (UL – LL)/3.92 by parametric method, and then, those analytes with a ratio greater than 0.4 were marked as "P < NP" (or "NP < P"). The (UL − LL)/3.92 term represents the standard deviation (SDri) in calculating the RI. The SD corresponds to gross between-individual SD (SD_G) containing within-individual SD (SD_I): $\sqrt{(SD_G \text{ ^2} + SD_I \text{ ^2})}$. From Fraser's theory of "allowable bias" in laboratory tests, desirable bias limit in laboratory tests is 0.25 × SDri, and minimum bias limit is 0.375 × SDri. We chose 0.4 after rounding up the value of 0.375 and used the cutoff in judging obvious between-method difference (bias).

Protocol for Transforming and Comparing Measurements Collected at Different Medical Centers

The protocol recommended by the IFCC C-RIDL for transforming healthy participant measurements from a different medical center involves measuring the same panel of reference serum samples from healthy individuals at both laboratories.[3] In accordance with CLSI guideline (EP9-A2),[14] *Measurement Procedure Comparison and Bias Estimation Using Patient Samples*, 40 samples were included in the comparison panel. Cross-comparison plots were made for each analyte with the ARUP results on the horizontal axis (x) and Mayo Clinic on the vertical axis (y). Linear regression lines were established using reduced major axis regression. The coefficient of variation (CV) of slope b was calculated as $CV\ (b)=100\sqrt{\frac{1-r^2}{n-2}}$ where r is the linear correlation coefficient. The optimal level for CV(b) is 5.5%,[3] and this level was selected for deciding when to allow conversion of the reference values

DOI: 10.1093/ajcp/aqy082

(RVs). The distributions of the analyte results from the ARUP populations were compared to the distributions from the Mayo populations using Mann-Whitney tests. Similarly, the transformed ARUP distributions based on the regression comparisons of the reference samples were compared with the original Mayo population using Mann-Whitney tests.

Results

Consistency of RIs Among the Six Calculation Procedures

The "optimal limits" for analytic bias are listed in the second column of Table 1. The ratios of the reference limit differences to the index method reference range were tabulated for the 15 pairwise combinations for each of the 28 analytes. These ratios were compared with the "optimal" analytic bias limits, for both the lower and upper reference range limits for both men and women. The counts of the number of these comparisons exceeding the "optimal" bias limits (maximum of 4) are tabulated in Table 1. Only two analytes (blood urea nitrogen [BUN] and iron) of the 28 analytes had no significant differences for the lower and upper limits for both women and men. The analytes with the largest number of divergent limits were potassium (49), calcium (45), sodium (44), albumin (43), and magnesium (42). These analytes have relatively tight "optimal" bias limits of 0.90%, 0.41%, 0.12%, 0.72%, and 0.90%, respectively, which probably contribute to these flags for divergent limits. The best agreement between reference limits was found when comparing the Mayo-SAS

Table 1
Tabulation of the Number of the Four Reference Limits (Male and Female High and Low Limits) Having Different Reference Limits Recommendations[a]

		Calculation Method															
		Mayo-SAS Quintile					EP Evaluator				LAVE(–)NP			LAVE(–)P		LAVE(+) NP	
Analyte	Optimal Bias Limits, %	EP Evaluator	NP(–)	P(–)	NP(+)	P(+)	NP(–)	P(–)	NP(+)	P(+)	P(–)	NP(+)	P(+)	NP(+)	P(+)	P(+)	Sum
Albumin	0.72	4	2	3	1	3	2	2	4	3	3	3	4	4	2	3	43
ALP	3.4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ALT	5.7	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	3
Amylase	3.7	1	1	1	2	2	0	0	1	1	0	2	2	1	0	1	15
AST	3.3	1	0	2	2	2	1	1	1	2	2	2	2	1	2	1	22
BUN	2.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Calcium	0.41	1	2	3	3	4	2	4	3	4	4	4	4	1	3	3	45
CEA	7.1	0	0	1	1	1	0	2	1	1	1	1	1	1	0	1	12
Cholesterol	2.1	0	1	1	0	1	1	0	1	2	0	1	1	1	1	0	11
CK	5.8	0	0	1	0	0	1	0	1	0	1	0	1	1	0	1	7
Chloride	0.25	0	1	4	0	4	1	4	0	4	4	1	4	4	3	4	38
Cortisol	5.1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
GGT	5.5	1	0	2	2	2	1	1	1	1	1	1	2	1	2	2	20
Glucose	1.2	1	1	2	4	3	1	1	4	2	2	4	2	3	1	1	32
HDL	2.8	0	0	1	0	1	0	0	0	1	1	0	1	1	0	1	7
Iron	4.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LDH	4.5	2	0	0	0	0	1	1	2	1	0	1	0	0	0	0	8
Magnesium	0.9	1	1	4	2	4	2	4	3	4	3	2	4	3	2	3	42
Phosphorus	1.7	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	2
Potassium	0.9	2	4	3	3	4	3	4	4	4	4	3	3	2	3	3	49
PRL	5.2	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	4
PTH	4.4	1	0	1	0	1	1	0	0	0	1	0	1	1	0	0	7
Sodium	0.12	1	1	4	1	4	2	4	2	4	4	1	4	4	4	4	44
Total bilirubin	4.5	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	3
Total protein	0.68	2	1	3	1	3	1	1	4	3	2	3	3	3	2	3	35
Triglycerides	4.8	0	0	0	0	0	0	0	2	0	0	0	1	0	0	1	4
TSH	3.9	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	11
Uric acid	2.8	0	0	0	1	0	0	0	1	1	0	1	0	1	0	0	5
Total (n = 240)		19	15	37	25	42	20	32	36	41	34	32	43	36	25	34	

ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CEA, carcinoembryonic antigen; CK, creatine kinase; GGT, γ-glutamyltransferase; HDL, high-density lipoprotein; LAVE, latent abnormal values exclusion; LDH, lactate dehydrogenase; NP, nonparametric; P, parametric; PRL, prolactin; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone; +, positive; –, negative.
[a]Pairwise comparisons were made between the six methods for calculating reference limits. These data are based on the 240 healthy participants from the Mayo Clinic study. The bottom row shows the total number of reference limits exceeding optimal bias by pairwise groups.

DOI: 10.1093/ajcp/aqy082

quintile vs the IFCC nonparametric method without LAVE exclusions, LAVE(–)NP, with 15 discrepant limits. The second best agreement in reference limit comparisons was found when comparing the Mayo-SAS quintile vs the EP Evaluator, with 19 discrepant limits. The two parametric procedures, LAVE(+)P and LAVE(–)P, agreed reasonably well with 25 discrepant limits.

Comparison of the Widths of the RIs Generated by the Six Procedures

With the small data set (n = 240), no appreciable between-method differences were observed for 11 (39%) analytes: albumin, BUN, high-density lipoprotein, sodium, calcium, potassium, chloride, iron, ALP, PRL, and PTH. However, for the last six analytes, 90% confidence intervals (CIs) by nonparametric methods were wider at either or both limits of the RIs. In Table 2, the ratio of the width of the 90% CI by a given method to that by the LAVE(+)P method exceeding 1.5 has been bolded. By reference to the actual data points in the corresponding scatterplot, which shows that the shape of their central part was close to Gaussian, it is obvious that the wider 90% CI was caused by extreme values unmatched to those in the central part (ie, the nonparametric method is susceptible to their presence, while the parametric method avoids their influence by including the truncation step after Gaussian transformation as described in the Materials and Methods).

In other analytes, those extreme values not only expanded 90% CIs but also raised the ULs calculated by nonparametric methods for glucose, AST, ALT, GGT, and CEA in both sexes; for CK, LDH, and TSH in men; and for uric acid, total bilirubin, and amylase in women. On the other hand, extreme values on the lower side caused lowering of the LL for cholesterol, magnesium, and cortisol in women. In tabulating Table 2, a critical value equivalent to UL ratio (or LL ratio) greater than 0.4 is marked as "P < NP."

By comparing ULs by LAVE(–)P and LAVE(+)P, the effect of the LAVE method was observed in RIs for

Table 2
Illustration of How Calculation Methods Affect Reference Intervals[a]

		Mayo (n = 240)				Mayo + ARUP (n = 490)			
Analyte	**Skew**	**LL-M**	**LL-F**	**UL-M**	**UL-F**	**LL-M**	**LL-F**	**UL-M**	**UL-F**
Albumin	–0.25	○	○	○	○	○	○	○	○
ALP	1.11	○	○	○	○	○	○	**P < NP**	**P < NP**
ALT	3.22	NP < P	○	**LAVE; P < NP**	**LAVE; P < NP**	○	○	**LAVE; P < NP**	**LAVE; P < NP**
Amylase	1.83	○	○	○	**LAVE; P < NP**	○	○	**P < NP**	**LAVE; P < NP**
AST	6.30	○	○	**LAVE; P < NP**	**P < NP**	○	○	**LAVE; P < NP**	○
BUN	0.58	○	○	○	○	○	○	○	○
Calcium	–0.39	○	○	○	○	○	○	○	○
CEA	5.80	○	○	**P < NP**	**P < NP**	○	○	○	○
Chloride	–0.60	○	○	○	○	○	○	○	○
Cholesterol	1.15	○	NP < P	**LAVE**	**P < NP**	○	NP < P	**P < NP**	**P < NP**
CK	4.62	○	○	**P < NP**	○	○	NP < P	**P < NP**	P < NP
Cortisol	0.77	○	**NP < P**	○	○	○	○	○	○
GGT	3.73	○	○	**LAVE; P < NP**	**LAVE; P < NP**	○	NP < P	**LAVE; P < NP**	P < NP
Glucose	1.31	○	○	**P < NP**	**P < NP**	○	○	**P < NP**	**P < NP**
HDL	0.66	○	○	○	○	○	NP < P	○	○
Iron	1.27	○	○	○	○	NP < P	○	**NP < P**	NP < P
LDH	6.70	○	○	**P < NP**	○	○	○	**P < NP**	○
Magnesium	–0.45	○	**NP < P**	○	○	○	○	○	○
Phosphorus	0.43	○	○	○	○	○	○	○	○
Potassium	1.38	○	○	○	○	○	○	○	○
PRL	3.98	○	○	○	○	○	○	○	○
PTH	1.11	○	○	○	○	○	○	○	○
Sodium	–1.73	○	○	○	○	**NP < P**	**NP < P**	○	○
Total bilirubin	2.08	○	○	○	**P < NP**	○	○	**P < NP**	**P < NP**
Total protein	–0.03	○	○	○	P < NP	○	○	○	○
Triglycerides	2.14	○	○	LAVE	○	○	○	○	○
TSH	8.21	○	○	**P < NP**	○	○	○	○	○
Uric acid	0.57	○	○	○	P < NP	○	○	P < NP	P < NP

ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CEA, carcinoembryonic antigen; CK, creatine kinase; GGT, γ-glutamyltransferase; HDL, high-density lipoprotein; LAVE, latent abnormal values exclusion lowered upper limit; LDH, lactate dehydrogenase; LL-F, lower limit of the reference interval for females; LL-M, lower limit of the reference interval for males; NP < P, lower limit by nonparametric methods is lower than parametric method; P < NP, upper limit by nonparametric methods is higher than parametric method; PRL, prolactin; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone; UL-F, upper limit of the reference interval for females; UL-M, upper limit of the reference interval for males; ○, reference intervals of all methods are similar.
[a]Bold values indicate that the 90% confidence intervals by nonparametric methods are wider than parametric method.

DOI: 10.1093/ajcp/aqy082

triglycerides, cholesterol, AST, ALT, and amylase, whose RVs had close associations with those of reference tests described in the Materials and Methods. In Table 2, the same scheme as just described above was used for judging the significance of difference in UL (or LL) between LAVE(–)P and LAVE(+)P methods.

It is notable that the occurrence of extreme values mainly depends on the skewness of the distribution shown in the second column of Table 2. A high absolute value for the skewness indicates either the distribution itself is unsymmetrical or the presence of inappropriate values caused by metabolic syndrome, noncompliance to requirements for food, and exercise prior to the sampling.

It is of note that the EP Evaluator method is primarily based on the CLSI guideline and calculates the RI by the nonparametric method. However, if the sample size is below 120, it computes the RI parametrically based on the classical Box-Cox formula. Therefore, it computed the RI parametrically for men with n = 92 and nonparametrically for women with n = 148. This distinction is well reflected in the graphs. In men, the width of the 90% CI by the EP Evaluator is closer to that by LAVE(–)P and LAVE(+)P, since the EP Evaluator uses a parametric method when there are fewer than 120 data values. While in women, its 90% CI is quite similar to those of the Mayo-SAS quantile and LAVE(–)NP methods.

In contrast, by use of the merged data set (n = 490) shown in the lower panels of the bar chart, the width of the 90% CI was conspicuously decreased in most analytes. The only exceptions were sodium, magnesium, and PRL. Their 90% CIs became rather broader by inclusion of more extreme values in the merged data. No obvious between-method differences were observed for 14 (48%) analytes: total protein, albumin, BUN, triglycerides, potassium, chloride, phosphorus, PTH, calcium, magnesium, CEA, cortisol, PRL, and TSH. However, 90% CIs were wider by nonparametric methods for the last six analytes due to the presence of extreme values, which were plotted outside of the graph frame.

Raised levels of ULs by nonparametric-based methods, attributable to multiple unmatched values in the periphery, are apparent for uric acid, total bilirubin, glucose, cholesterol, iron, AST, ALT, LDH, ALP, GGT, CK, and amylase. The effect of LAVE to lower the UL was noted in AST, ALT, GGT, and CEA. The data size after the LAVE procedure was reduced by about 12% to 14%; however, no appreciable increase in the width of 90% CIs was observed.

Limitations of Linear Transformations to Improve Homogeneity of Merged Data

The comparison of the reference samples showed that all but six of the analytes had CV(*b*) of the regression slopes less than the 5.5% optimal level. Table 3 shows that the high CV(*b*) analytes (in bold) were calcium, CEA, chloride, magnesium, sodium, and total protein. In comparing the ARUP and Mayo Clinic original frequency distributions of the test values for the 28 analytes by sex, 10 analytes differed by a *P* value of .01 or less for men, and 12 analytes were divergent for women. After transformation of the ARUP test values, the frequency distributions for all but three of the analytes for men were corrected. However, for women, eight of the 12 divergent analytes were still divergent after correction. It is notable that six of the 18 nondivergent analytes for men became divergent after transformation and four of the 16 originally nondivergent analytes became divergent for women. When the analytic methods are similar, the small analytic differences may be obscured by the variability of the regression statistics. Also, the linear transformation protocol is only intended to correct for analytic measurement differences. Any population-based differences or preanalytic differences would remain. Because of these limitations, the ARUP test value data were merged with the Mayo Clinic data without transformations. These merged data were used to analyze the effects of a larger data set for estimating RIs.

RIs

The RIs and their confidence limits were determined separately for each sex using each of the six procedures with two data sets: one from Mayo's results (n = 240) and the other from merged results of Mayo and ARUP (n = 490). The comparison graphs for all 28 analytes are shown in Supplemental Figure 1 (all supplemental materials can be found at *American Journal of Clinical Pathology* online) and for six representative analytes in Figure 1. It is important to note that for analytes with skewed distribution, the scale of the x-axis was power transformed to make the distribution of RVs close to the Gaussian shape. Power was set to 0.4 for ALP; 0.3 for total bilirubin, AST, LDH, amylase, cortisol, PTH, and TSH; and 0.2 for triglycerides, ALT, GGT, CK, CEA, and PRL. These values of power were set in reference to the results of the parametric method applied to each analyte. For other analytes, no transformation was made by setting power = 1.0.

Discussion

No gold-standard protocol exists for defining RIs. Most investigators agree that an unbiased estimate of the central 95% range of values in healthy individuals is recommended. However, the definition of *healthy* can vary substantially. No agreed-upon criteria exist for the elimination of inappropriate extreme values. Large sample sizes

DOI: 10.1093/ajcp/aqy082

Table 3
Comparison of Reference Samples and the Frequency Distributions of Test Values From Two Centers Before and After Transformation[a]

	Linear Regression Coefficients of Reference Test Panel Results			Comparison of Original Frequency Distributions: ARUP and Mayo Clinic		Comparison of Distributions: Transformed ARUP vs Mayo Clinic	
Analyte	CV(*b*)	Slope	Intercept	Men	Women	Men	Women
Albumin[b]	5.42	0.952	0.213	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
ALP[b]	1.44	0.972	2.941	0.1148	0.0574	0.0477	0.0139
ALT[b]	2.27	1.116	0.166	**<0.0001**	**<0.0001**	0.0298	0.0469
Amylase[b]	1.3	1.016	0.426	0.2999	0.63	0.5967	0.2603
AST[b]	3.96	1.108	–0.293	0.1806	0.1716	0.4913	0.1868
BUN[b]	3.9	0.954	0.042	0.9965	0.0765	0.1376	0.6018
Calcium[b]	**8.75**	0.858	1.423	0.0372	0.0358	**<0.0001**	**<0.0001**
CEA	**6.08**	0.922	–0.017	0.1796	0.5272	0.9402	0.0788
Cholesterol[b]	2.23	0.981	5.506	0.3754	0.108	0.1675	0.0309
CK[b]	0.58	1.02	0.005	0.2819	**0.0085**	0.4146	0.0244
Chloride[b]	**9.88**	1.082	–6.213	0.4162	0.9981	**<0.0001**	**<0.0001**
Cortisol	3.15	0.72	0.808	**<0.0001**	**<0.0001**	0.1668	0.7363
GGT[b]	1.41	0.955	0.191	0.1219	0.1543	0.274	0.4832
Glucose[b]	1.9	0.97	1.769	0.003	0.0043	0.0268	0.0302
HDL[b]	2.13	1.024	0.608	0.0348	0.0141	0.0672	0.0459
Iron[b]	0.88	1.019	2.752	0.2964	0.0142	0.9402	**0.0003**
LDH[b]	2.57	1.005	–1.615	0.0391	0.9883	**0.0004**	0.0754
Magnesium[b]	**6.63**	1.803	–0.065	**0.0004**	**0.0002**	**<0.0001**	**<0.0001**
Phosphorus[b]	3.37	0.924	0.139	**0.0013**	0.0558	0.2431	0.5966
Potassium[b]	2.27	1.011	0.097	**<0.0001**	**<0.0001**	0.4436	**0.0072**
PRL	1.62	0.987	0.285	**<0.0001**	**0.0002**	**<0.0001**	**0.0006**
PTH	2.61	1.027	3.968	0.1123	0.393	0.9372	0.6383
Sodium[b]	**13.1**	0.941	10.916	**<0.0001**	**<0.0001**	**0.0001**	**0.0002**
Total bilirubin[b]	2.7	0.976	0.033	**0.0009**	**<0.0001**	**0.0346**	**0.0033**
Total protein[b]	**6.65**	0.951	0.532	0.9034	0.0266	**0.0001**	**<0.0001**
Triglycerides[b]	0.79	0.994	0.036	0.042	**0.0013**	0.0515	**0.0019**
TSH	0.73	1.008	0.023	0.8939	0.4107	0.716	0.2724
Uric acid[b]	0.95	1.006	0.0116	0.0108	**<0.0001**	**0.0038**	**<0.0001**

ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CEA, carcinoembryonic antigen; CK, creatine kinase; CV, coefficient of variation; GGT, γ-glutamyltransferase; HDL, high-density lipoprotein; LAVE, latent abnormal values exclusion; LDH, lactate dehydrogenase; NP, nonparametric; P, parametric; PRL, prolactin; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone.
[a]Linear regression of reference samples measured at ARUP (x) and Mayo Clinic (y). Comparison of the frequency distributions of the original test values collected and measured at ARUP vs values for samples collected and measured at Mayo Clinic. The last columns compare the frequency distributions of the transformed ARUP values with the original Mayo Clinic values. Differences were evaluated with Mann-Whitney tests. Bold values represent coefficient of slope, CV(*b*), values exceeding optimal level of 5%, and highly significant *P* values.
[b]Analytes were measured by same method, Roche Modular, at both sites.

are needed for robust nonparametric estimates. The CLSI guideline recommends using a minimum data set of 120 healthy participants for establishing RIs for each important subgroup, such as sex and race.[1] However, RIs derived from data sets that small often have larger confidence intervals compared with those derived from data sets 200 or larger. In the absence of a gold standard, we evaluated six statistical procedures for calculating RIs in terms of two criteria: (1) consistency of assigned reference limits and (2) width of the RIs. Unfortunately, although the nonparametric procedures generated similar RIs, they also were most subject to interference by extreme values and generated RIs they may have inappropriately wide intervals.

The nonparametric methods are generally regarded as the most robust procedures for providing unbiased estimates of the central 95% RIs. However, these procedures may fail when the number of inappropriate values exceeds the capacity of the conventional outlier exclusion method such as the Dixon or Tukey method, which is applicable only for cases with one or few outliers. In conducting a study for determining RIs, it is practically impossible to totally avoid inclusion of inappropriate values caused by the presence of a sizable number of individuals with latent disorders of high prevalence like metabolic syndrome or inflammation or anemia, or individuals who did not follow precaution on food and muscular exertion.

On the other hand, the parametric method after Gaussian transformation of the RV distribution can reduce the influence of those values by truncating once at mean ± kSD (the C-RIDL procedure uses k = 2.57, which corresponds to exclusion of 1% of RVs). The reason why the CLSI/IFCC guideline does not recommend the parametric

DOI: 10.1093/ajcp/aqy082

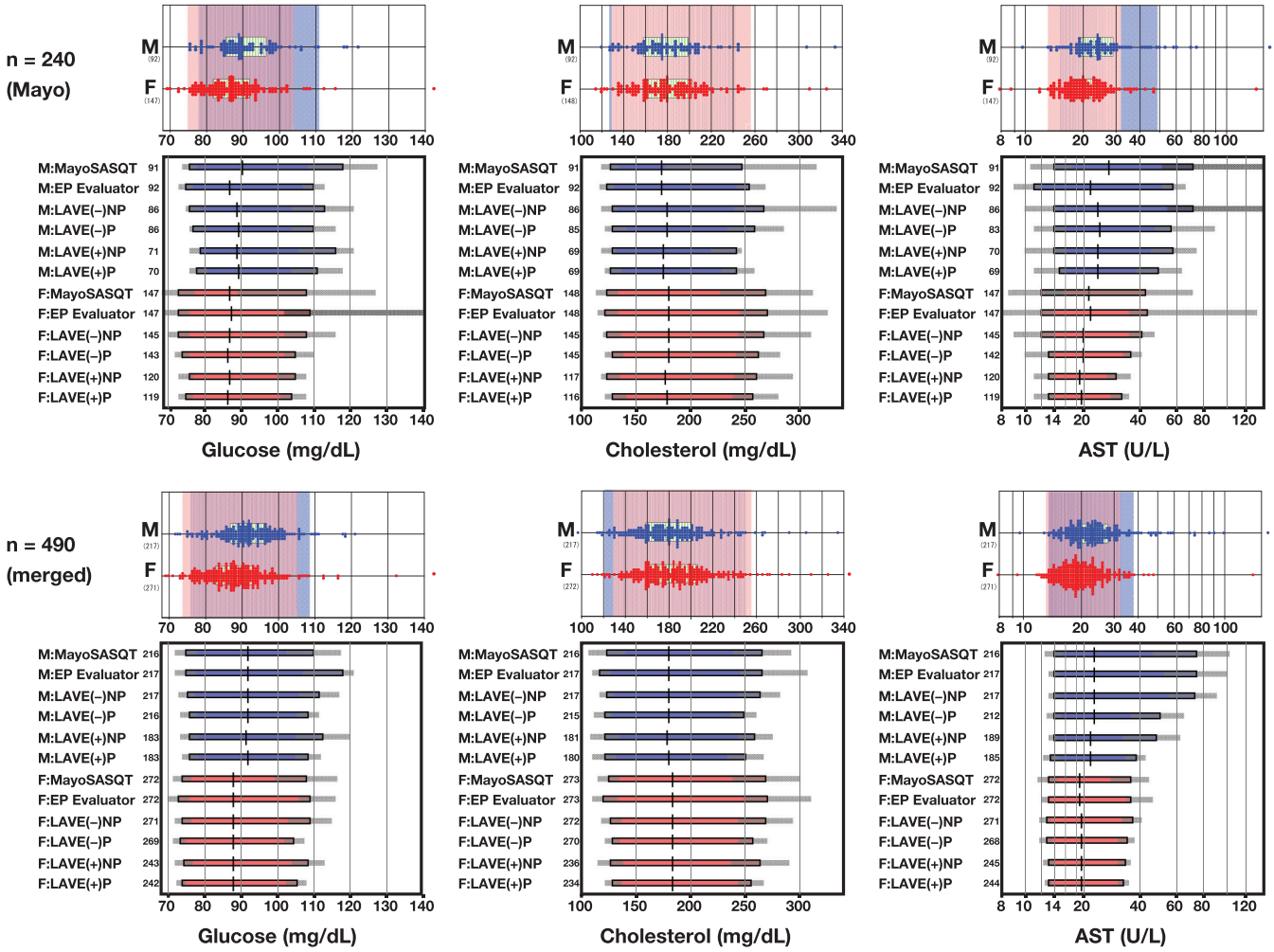

Figure 1 Shown are bar chart comparisons of reference intervals (RIs) derived for glucose, cholesterol, aspartate aminotransferase (AST), γ-glutamyltransferase (GGT), creatine kinase (CK), and thyroid-stimulating hormone (TSH), calculated by six methods: Mayo SAS Quantile, EP Evaluator, and four International Federation of Clinical Chemistry and Laboratory Medicine methods: parametric (P) and nonparametric (NP) without exclusion (LAVE(–)) and with exclusion (LAVE(+)). The upper panels in each set are for RIs derived from 240 Mayo Clinic health study participants. The lower panels in each set are for RIs derived from merged ARUP and Mayo Clinic data composed of 490 participants. Red bar is for females (F) and blue is for males (M).

method was unreliability of the classic Box-Cox power transformation formula to achieve the Gaussian shape.[2] However, the modified Box-Cox formula,[10] which includes a parameter representing the origin of transformation, was proved invariably successful to achieve Gaussian shape if the distribution is unimodal and there are no results below detection limits.[8,15] Presumed weakness of the parametric method is increased uncertainty in estimating the RI when the sample size is small with n < 120, because the method is too flexible to include extreme values in the periphery. Despite this concern, actual performance of the parametric method with respect to 90% CIs using the small RVs from 92 men and 148 women was in general equivalent to the nonparametric methods and sometimes much better when extreme values were present. In any case, with increment of data size, the parametric method reflects data points in the central part of the distribution more closely and gets less affected by extreme values in the periphery with a data size of 200 or more.[15]

These findings clearly indicate that the parametric method is recommended when a certain number of inappropriate values unmatched to the central profile of the RVs are expected. If those inappropriate values are sizable in number due to inclusion of a group of individuals with common disorders like metabolic syndrome, their influence cannot be removed univariately. The use of the LAVE procedure, although it is regarded as an empirical, not definitive, method, was proved to reduce the influence of such a category of inappropriate values in the derivation of the RIs.[9,15]

DOI: 10.1093/ajcp/aqy082

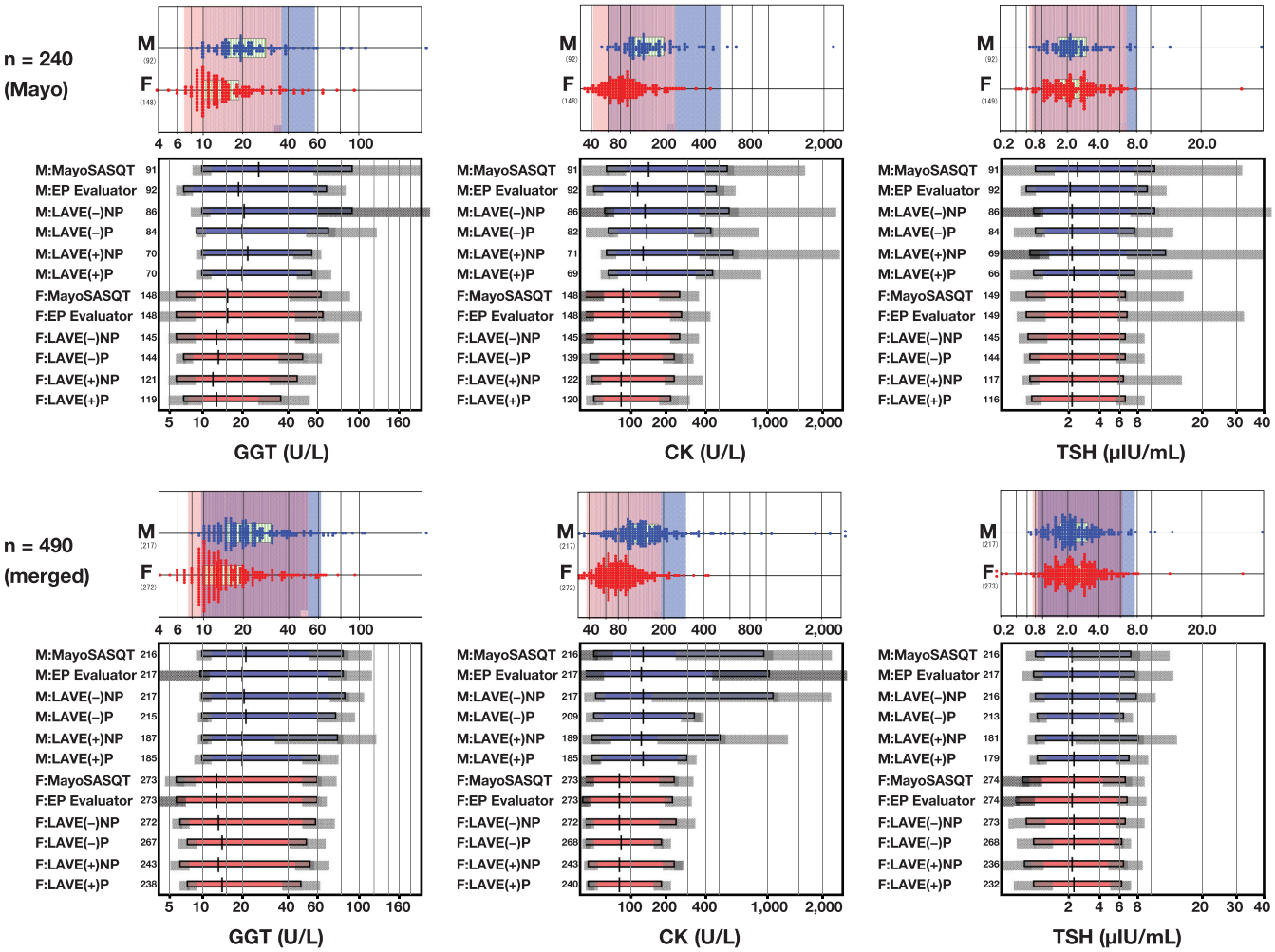

Figure 1 (cont) Shaded bars in gray represent 90% confidence limits. The number shown beside the method names indicates data size used for the calculation. Scatterplots on top of each bar chart represent actual sex-specific distributions of reference values (RVs). Out-of-scale values were plotted just outside of the graph frame. Blue and red shades represent RIs for M and F, respectively, derived by the LAVE(+)P method. For analytes with skewed distribution, the scale of the x-axis was power transformed to make the distribution of RVs close to the Gaussian shape (see the main text for the value of power used for transformation).

This study compared analytic methods that were mainly performed on the same commercial measurement system (Roche Modular). Therefore, the analytic differences generally were small. In comparing laboratories with different analytic instruments, larger differences would be expected. The transference algorithm based on comparing the analytic differences found when measuring reference samples probably would work better for correcting the RIs when larger differences occur.

The use of merged data sets helped harmonize the results of the various calculation methods and provides more unbiased estimates of the RIs.

Our evaluation of the procedure proposed by Ichihara et al[3,16] for transforming RVs based on comparison of test results for a set of sera measures in common (a subset of volunteers' serum samples[16] or the serum panel specifically made for the comparison[3]) shows this method generally works but does not work universally. Our study confirmed that a high CV(*b*) in the reference samples limits the utility of this method.[3] None of the six analytes with high CV(*b*) showed any significant changes in the divergence of the patient distributions after transformation. It is important to note that automatic transformation of distributions that are statistically similar may result in more divergences. A take-home message from this study is that linear regression–based conversion of values should be done only when notable bias is observed between the values of the two centers.

Another unambiguous message obtained from this study is the importance of data size for deriving the

DOI: 10.1093/ajcp/aqy082

reproducible RIs. Although the minimum sample size is suggested as 120 by the CLSI guideline, our results indicate that a larger sample size is needed for multicenter studies intended for deriving RIs for common use.

Conclusions

The statistical procedure used to analyze the reference subject data can influence the RIs. Nonparametric procedures generally generate similar results, but the RIs may be inappropriately wide due to the inclusion of inappropriate values. Since it is difficult to eliminate participants with inappropriate extreme values, parametric methods are recommended. Large data sets are recommended for robust estimates. The CLSI guideline of at least 120 participants often is too small for reliable estimates. We recommend at least 200 participants for each subgroup. Merging data from multiple institutions is a potential method of obtaining larger data sets. However, one should be cautious when merging data because analytic regression transformations may not correct for all intrainstitutional differences.

Corresponding author: George G. Klee, MD, PhD, Mayo Foundation, 200 First St SW, Rochester, MN 55905; klee.george@mayo.edu.

References

1. Ichihara K, Ozarda Y, Barth JH, et al; Committee on Reference Intervals and Decision Limits, International Federation of Clinical Chemistry and Laboratory Medicine and Science Committee, Asia-Pacific Federation for Clinical Biochemistry. A global multicenter study on reference values: 2. Exploration of sources of variation across the countries. *Clin Chim Acta*. 2017;467:83-97.
2. Horowitz GL, Altaie S, Boyd JC, et al. *CLSI EP28-A3C:2010 Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline–Third Edition*. Wayne, PA: Clinical and Laboratory Standards Institute; 2010.
3. Ichihara K, Ozarda Y, Klee G, et al; Committee on Reference Intervals and Decision Limits, International Federation for Clinical Chemistry and Laboratory Medicine. Utility of a panel of sera for the alignment of test results in the worldwide multicenter study on reference values. *Clin Chem Lab Med*. 2013;51:1007-1025.
4. Ozarda Y, Ichihara K, Barth JH, et al; Committee on Reference Intervals and Decision Limits (C-RIDL), International Federation for Clinical Chemistry and Laboratory Medicine. Protocol and standard operating procedures for common use in a worldwide multicenter study on reference values. *Clin Chem Lab Med*. 2013;51:1027-1040.
5. Chen LC. An introduction to quantile regression and QUANTREG procedure. http://www2.sas.com/proceedings/sugi30/213-30.pdf. Accessed July 12, 2017.
6. Essermeant L. Normal ranges determination with quantile regression. http://ncs-conference.org/download/slide/20120926_Wednesday/1330_1530_various/4_Essermeant.pdf. Assessed July 12, 2017.
7. Koenker R, Hallock K. Quantile regression: an introduction. *J Econ Perspect*. 2001;15:143-156.
8. Data Innovations. EP Evaluator, quality assurance . . . simplified. http://datainnovations.com/ep-evaluator. Accessed July 12, 2017.
9. Ichihara K, Ozarda Y, Barth JH, et al; Committee on Reference Intervals and Decision Limits, International Federation of Clinical Chemistry and Laboratory Medicine. A global multicenter study on reference values: 1. Assessment of methods for derivation and comparison of reference intervals. *Clin Chim Acta*. 2017;467:70-82.
10. Ichihara K, Boyd JC; 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*. 2010;48:1537-1551.
11. Ozarda Y, Ichihara K, Bakan E, et al. A nationwide multicentre study in Turkey for establishing reference intervals of haematological parameters with novel use of a panel of whole blood. *Biochem Med*. 2017;27:350-377.
12. Ricos C, Alvarex V, Cava F, et al. Optimal Biological Variation database specifications. https://www.westgard.com/optimal-biodatabase1htm.htm. Accessed July 12, 2017.
13. Ricos C, Alvarez V, Cava F, et al. Desirable specifications for total error, imprecision, and bias, derived from intra- and inter-individual biologic variation. https://www.westgard.com/biodatabase1.htm. Accessed July 12, 2017.
14. Budd JR, Durham AP, Gwise TE, et al. *Measurement Procedure Comparison and Bias Estimation Using Patient Samples, Clinical and Laboratory Standards Institute Approved Guideline*. Vol 30. Wayne, PA: Clinical and Laboratory Standards Institute; 2013.
15. Ichihara K. Statistical considerations for harmonization of the global multicenter study on reference values. *Clin Chim Acta*. 2014;432:108-118.
16. Ichihara K, Ceriotti F, Kazuo M, et al; Committee on Reference Intervals and Decision Limits, International Federation for Clinical Chemistry and Laboratory Medicine, and the Science Committee for the Asia-Pacific Federation of Clinical Biochemistry. The Asian project for collaborative derivation of reference intervals: (2) results of non-standardized analytes and transference of reference intervals to the participating laboratories on the basis of cross-comparison of test results. *Clin Chem Lab Med*. 2013;51:1443-1457.

DOI: 10.1093/ajcp/aqy082