Clinica Chimica Acta 520 (2021) 186–195

Contents lists available at ScienceDirect

Clinica Chimica Acta

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

Comparison of reference intervals derived by direct and indirect methods based on compatible datasets obtained in Turkey

Yesim Ozarda[a,*], Kiyoshi Ichihara[b], Graham Jones[c,d], Thomas Streichert[e], Robab Ahmadian[f], on behalf of the IFCC Committee on Reference Intervals and Decision Limits (C-RIDL)

[a] Department of Medical Biochemistry, Istanbul Health and Technology University School of Medicine, Istanbul, Turkey
[b] Faculty of Health Sciences, Yamaguchi University Graduate School of Medicine, Ube, Japan
[c] Department of Chemical Pathology, SydPath, St Vincent's Hospital, Sydney, NSW, Australia
[d] University of NSW, Sydney, NSW, Australia
[e] Institute for Clinical Chemistry, Faculty of Medicine, University of Cologne, Cologne, Germany
[f] Department of Statistics, Uludag University School of Medicine, Bursa, Turkey

ARTICLE INFO

Keywords:
Laboratory information system (LIS)
Power transformation
Parametric method
Hoffmann method
Bhattacharya method
Arzideh method
Wosniok method
Latent abnormal values exclusion (LAVE)

ABSTRACT

Background: Indirect derivation of reference intervals (RIs) from the laboratory information system (LIS) has been recently pursued. We aimed at evaluating the accuracy of indirectly predicted RIs compared to the RIs established directly from healthy subjects in the nationwide RI study in Turkey, targeting 25 major chemistry analytes.

Methods: LIS data were retrieved from the laboratory that performed measurements for the direct study. They were cleaned by limiting to outpatients with age 18–65 years, and by allowing only one record per year per patient. Evaluated were four indirect methods of univariate approach: Hoffmann, Bhattacharya, Arzideh, and Wosniok methods. Power transformation of the LIS dataset was performed either using the power (λ) reported by the IFCC global RI study (the first two methods) or using a λ predicted (the last two).

Results: Compared to the direct study dataset, the LIS dataset showed a variable degree of alterations in peak location and shape. Consequently, lower-side peak-shifts observed in sodium, albumin, etc. led to lowered RI limits, whereas higher-side peak-shift observed in triglyceride, low-density lipoprotein cholesterol, etc. led to raised RI limits. Overall, 72% (62–81) of the RI limits predicted by indirect methods showed significant biases from direct RIs. However, the biases observed in total cholesterol, lactic dehydrogenase, etc. were attributed to a higher-side age-bias in LIS dataset. After excluding them, the overall proportion of biased RIs was reduced to 47% (38–54).

Conclusion: To reduce prediction biases that remained after age adjustment, it is necessary to apply more rigorous data-cleaning before applying indirect methods.

1. Introduction

The recommended process for producing reference intervals (RI) is the direct method with a priori selection of members of the reference population as described in the international guideline issued by the Clinical and Laboratory Standards Institute (CLSI) [1]. This process, however, is extremely difficult to carry out properly. For reproducible results, it requires recruitment of a large number of well-defined healthy individuals, ideally by conducting a multicenter study [2] as well as specification of proper schemes for specimen collection and handling,

Abbreviations: Alb, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AMY, amylase; AoMN, average of multivariately normals; BR, bias ratio; Ca, calcium; CDL, clinical decision limit; CK, creatine kinase; CI, confidence interval; Cl, chloride; CLSI, Clinical and Laboratory Standards Institute; Cre, creatinine; DBil, direct bilirubin; GGT, gamma-glutamyltransferase; Glu, glucose; HDL-C, HDL-Cholesterol; IP, inorganic phosphate; K, potassium; LAVE, latent abnormal values exclusion; LDH, lactate dehydrogenase; LDL-C, LDL-Cholesterol; LIS, laboratory information system; LL, lower limit; Me, median; Mg, magnesium; MN, multivariately normal; Na, sodium; NP, non-parametric; P, parametric; PN, power normal; RI, reference interval; RV, reference value; SD, standard deviation; TBil, total bilirubin; TC, total cholesterol; TG, triglycerides; TP, total protein; TMC, truncated minimum chi-square method; TML, truncated maximum likelihood method; UA, uric acid; UL, upper limit.

* Corresponding author at: Department of Medical Biochemistry, Istanbul Health and Technology University School of Medicine, Istanbul, Turkey.
E-mail address: yesim.ozarda@istun.edu.tr (Y. Ozarda).

https://doi.org/10.1016/j.cca.2021.05.030
Received 2 February 2021; Received in revised form 30 April 2021; Accepted 26 May 2021
Available online 1 June 2021

measurement, and comprehensive source-of-variation analyses and exclusion of inappropriate results before final determination of RIs. These processes are very time-consuming, and there are major costs incurred for every step of the process. Ethical clearance of the study may also consume substantial time. For smooth implementation of the study, significant efforts are required to obtain collaboration among researchers and any supportive diagnostic companies. Consequently, there have been many alternative proposals to make use of routine laboratory data stored in the laboratory information system (LIS) for deriving RIs, so called a posteriori or indirect method [3].

An early indirect method was described by Hoffman [4] in 1963, when computer-based data analysis is in its primitive stage without much theoretical knowledge about distribution of laboratory data. Hoffman observed that the distribution of routine test results, regardless of the analyte, has a central smooth-looking peak, which could be assumed to represent "normal" values and approximate a Gaussian distribution. Hence, he proposed to use a probability paper plot to dissect the peak and by manual linear regression of its central segment, and then to determine the limits of RI by extrapolation of the line. A problem of this method was that the assumption was always Gaussian, without considering other distribution patterns. In 1967, Bhattacharya [5] developed another graphical method to identify one or more Gaussian peaks in the histogram of observed data. This method has been applied to laboratory data assigning the largest peak to represent the reference population and to derive RIs. However, with improved knowledge on the distribution of reference values (RVs), transformation of the raw data to a Gaussian shape is acknowledged to be the essential procedure in applying the Hoffman and Bhattacharya methods [6,7].

Recently, Arzideh et al. [8] proposed to apply a power transformation of the source data by use of Box-Cox formula. Their method features an iterative algorithm to find the best truncation segment of RV distribution and to estimate the parameters of the corresponding distribution that best fits to the central "non-disease" segment via the maximum likelihood method, and thus named "truncated maximum likelihood" (TML) method. Truncated minimum chi-square (TMC) method was subsequently developed by Wosniok and Haeckel [9], which follows the same strategy as TML method except for an improved estimation of λ with more reliable testing for normality of the central truncated segment.

In recent years, the above-described methods of univariate approach are proposed for use in the indirect derivation of RIs [3]. However, the validation of those methods by simultaneous comparison of RIs with those by the a priori (direct) method has not been available so far. Another concern after reviewing the papers proposing those methods was inconsistencies in the efforts for pre-cleansing LIS source data to minimize unwanted over-representation of disease groups with overlapping results.

With these backgrounds, members of the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Committee on Reference Intervals and Decision Limits (C-RIDL) launched a project of scientifically validating each proposed method. The best scheme is to conduct a direct RI study ensuring sufficient sample size based on well-designed protocol, and to perform an indirect study in parallel using the LIS source data from the same clinical laboratory that provided assays for the direct study. Such circumstances were available in Turkey, which conducted a nationwide study for establishing RIs for chemistry analytes as a part of the global multicenter RI study [2]. No regional differences in reference values were observed in any analyte. Therefore, we retrieved a LIS dataset for the corresponding time period from the central laboratory and derived the RIs by four recently advocated indirect methods to evaluate their accuracy in predicting the RIs established by the direct study.

2. Methods

2.1. Source dataset for the direct method

The source data served for the direct methods were those obtained at the time of global multicenter RI study [13]. Blood samples were collected nationwide from 28 laboratories located in seven regions (≥400 samples/region, 3066 in all) in 2011–2012. We applied the inclusion and exclusion criteria described in the C-RIDL harmonized protocol [10]. The sex and age distributions for the target age of 18 to 65 were made nearly equal as shown in Table 1.

The sera were collectively measured for 25 analytes listed below in Uludag University in Bursa using Abbott reagents and analyzer (Architect 8000®, Abbott Diagnostics, IL, USA). By ANOVA, we confirmed no obvious between-region differences in test results.

Analytes comprised total protein (TP), albumin (Alb), urea, uric acid (UA), creatinine (Cre), total bilirubin (TBil), direct bilirubin (DBil), glucose (Glu), total cholesterol (TC), triglyceride (TG), high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), sodium (Na), potassium (K), chloride (Cl), calcium (Ca), inorganic phosphate (IP), magnesium (Mg), aspartate aminotransferase (AST), alanine aminotransferase (ALT), lactate dehydrogenase (LDH), alkaline phosphatase (ALP), gamma glutamyl transferase (GGT), creatine kinase (CK), amylase (AMY). For the assay, reference materials were measured for standardization of all the test results. Internal (Abbott Diagnostics, IL, USA) and External Quality Assurance Services (Bio-Rad EQAS) have been assessed to ensure the accuracy and stability of the test results.

2.2. Source dataset for indirect methods

We retrieved test results for 25 analytes measured, also by use of the Abbott analyzer, between 2011 and 2016 (6 years) from the LIS in the Clinical Laboratory of Uludag University Hospital. The age range of the source data was restricted to 18–65 years.

2.2.1. Preliminary data-cleaning

Then, we applied the following two measures of cleaning the data following the recommendation by Jones [3]. (1) After exclusion of test results of inpatients, only outpatient results were included except for those ordered from outpatient clinics specialized in emergency, oncology, anesthesiology and reanimation, gastroenterology, and nephrology. (2) If multiple records per year existed per patient, all records of the year were excluded except for single/first result, with an assumption that the necessity of multiple testing implies higher chances of non-healthy status. Stepwise changes in the data sizes following each exclusion procedure are shown in Table 2.

After this selection process, sex and age breakdown of the source data is shown for each analyte in **Suppl.** Table 1 and its total for all analytes combined are shown in Table 1.

2.2.2. Gaussian transformation of source data

In derivation of RIs by any indirect method or in applying Tukey outlier detection method, it is necessary to perform Gaussian transformation of source dataset as much as possible by use of Box-Cox power transformation: $X = (x^{\lambda} - 1)/\lambda$, where X represents transformed value of x using the power λ. In this collaborative study, to facilitate the search for appropriate λ, the λ values reported by C-RIDL [11] that are listed in **Suppl. Table 2** were made available for the methods that do not include any algorithm for estimation of parameters including λ. The reported λ values were obtained in the C-RIDL global multicenter study on reference values [11] as an analyte-by-analyte average of λs for 20 distributions (10 countries × 2 genders). For DBil, λ was not available and thus set to 0.25 that was used in the nationwide Turkish study [10].

Table 1
Sex and age distribution of source data used in the direct and indirect methods.

Age grp		18–29	30–39	40–49	50–59	60–65	Sum	Ave of age	SD of age
For direct methods									
Male	**n**	348	414	368	259	87	1476	39.6	11.6
	%	23.6	28.0	24.9	17.5	5.9	50.2		
Female	**n**	378	389	371	250	79	1467	38.8	11.8
	%	25.8	26.5	25.3	17.0	5.4	49.8		
Sum	**n**	726	803	739	509	166	2943	39.2	11.7
	%	24.7	27.3	25.1	17.3	5.6	100		
For indirect methods			Data after two-step exclusion						
Male	**n**	159,686	169,769	192,165	238,180	171,512	931,312	45.4	14.1
	%	17.1	18.2	20.6	25.6	18.4	36.5		
Female	**n**	242,408	310,911	364,901	447,322	255,450	1,620,992	47.5	12.7
	%	15.0	19.2	22.5	27.6	15.8	63.5		
Sum	**n**	402,094	480,680	557,066	685,502	426,962	2,552,304	46.6	13.3
	%	15.8	18.8	21.8	26.9	16.7	100		

Table 2
Changes in sizes of LIS source dataset of 25 analytes after two-step data cleaning.

Analyte	Changes in data size					Male-female ratio after cleaning			
	Total	Limited to outpatients	%	Limited to n = 1/patient	%	n (Male)	%	n (Female)	%
TP	701,789	380,108	54.2	66,144	9.4	24,788	37.5	41,356	62.5
Alb*	957,568	526,031	54.9	79,893	8.3	29,236	36.6	50,657	63.4
Urea	1,074,176	646,211	60.2	128,393	12.0	47,655	37.1	80,738	62.9
UA	1,797,796	1,060,311	59.0	170,727	9.5	62,062	36.4	108,665	63.6
Cre	719,926	385,845	53.6	81,697	11.3	29,020	35.5	52,677	64.5
TBil	1,800,268	1,085,770	60.3	223,798	12.4	80,758	36.1	143,040	63.9
DBil	838,006	450,255	53.7	57,954	6.9	21,607	37.3	36,347	62.7
Glu	793,421	427,421	53.9	55,683	7.0	20,605	37.0	35,078	63.0
TC	557,184	384,111	68.9	188,268	33.8	67,544	35.9	120,724	64.1
TG	439,287	295,817	67.3	99,150	22.6	37,963	38.3	61,187	61.7
HDL-C	432,031	291,589	67.5	97,764	22.6	37,482	38.3	60,282	61.7
LDL-C	395,781	264,831	66.9	89,427	22.6	34,124	38.2	55,303	61.8
Na	376,633	251,058	66.7	84,747	22.5	31,804	37.5	52,943	62.5
K	1,771,379	1,425,836	80.5	175,702	9.9	64,196	36.5	111,506	63.5
Cl*	1,797,393	1,059,911	59.0	170,698	9.5	61,890	36.3	108,808	63.7
Ca	1,707,571	1,004,155	58.8	158,036	9.3	54,187	34.3	103,849	65.7
IP	352,232	153,164	43.5	18,762	5.3	5,500	29.3	13,262	70.7
Mg*	440,190	215,561	49.0	16,825	3.8	5,332	31.7	11,493	68.3
AST	2,088,568	1,209,640	57.9	240,435	11.5	86,350	35.9	154,085	64.1
ALT	2,108,477	1,220,780	57.9	183,152	8.7	65,912	36.0	117,240	64.0
LDH*	443,989	229,236	51.6	41,226	9.3	15,469	37.5	25,757	62.5
ALP	209,681	129,080	61.6	55,860	26.6	19,749	35.4	36,111	64.6
GGT	263,368	175,430	66.6	40,321	15.3	16,925	42.0	23,396	58.0
CK	239,873	129,477	54.0	23,587	9.8	9,446	40.0	14,141	60.0
AMY*	103,174	58,338	56.5	4,055	3.9	1,708	42.1	2,347	57.9
Total	22,409,761	13,459,966	60.1	2,552,304	11.4	931,312	36.5	1,620,992	63.5

* Data of the analytes were not served for the method comparison study due to between-year bias in measurements.

2.3. Methods for derivation of RIs

2.3.1. Direct methods

2.3.1.1. Nonparametric method by CLSI: D-NP-CLSI. This method corresponds to the procedure recommended in the CLSI guideline [1]. The RI was derived, by nonparametric method, as the central 95% range of RVs. No measure for secondary exclusion was performed. Ninety % confidence interval (CI) was estimated by the bootstrap method through repeated resampling of 50 times. The final RI was set to the average of repeatedly derived LLs and ULs. This bootstrap procedure for 90% CI estimation was also applied to the other parametric methods.

2.3.1.2. Parametric method without LAVE: D-P-LAVE(-). Parametric method was based on modified Box-Cox formula with two parameters representing power (λ) and the origin of transformation (a) [12].

$$X = \frac{(x-a)^{\lambda} - 1}{\lambda}$$

where *x* and *X* represent test results before and after the transformation. Maximum likelihood method was used for first estimating ‘λ’, followed by estimation of ‘a’. This two-step estimation was performed iteratively until an optimal solution was obtained. Then, mean and SD under the transformed scale (m^T, SD^T) were computed after excluding values outside mean ± 2.81SD (0.005% of data in tails) once. RI limits (lower limit: LL; upper limit: UL) under the transformed scale ($LL^T \sim UL^T$) was reverse transformed to the original scale (LL ~ UL).

$$LL^T = m^T - 1.96SD^T$$

$$UL^T = m^T + 1.96SD^T$$

$$LL = (\lambda \times LL^T + 1)^{1/\lambda} + a$$

$$UL = (\lambda \times UL^T + 1)^{1/\lambda} + a$$

In this method, a preliminary procedure of secondary exclusion of abnormal results, called latent abnormal values exclusion (LAVE) method [11,12] described below, was not applied.

The RI derived by this method was set as a reference in measuring between-method bias of LL or UL, because of known problem of D-NP-CLSI regarding its susceptibility to outlying points and low precision (wider confidence intervals) of RI limits [11,13].

2.3.1.3. Parametric method with LAVE: D-P-LAVE(+). The same parametric method as above was used, but the LAVE procedure was applied in the iterative calculation to reduce the influence of common abnormal results attributable to highly prevalent nutritional abnormalities or to non-basal conditions prior to the sampling. The following eight nutritional/muscular markers were set as reference tests used for judging inappropriate results: Alb, UA, TG, AST, ALT, LDH, GGT, and CK. If any individual has two or more abnormal results among the reference tests other than the one under derivation of the RI, the value of that individual was excluded from the calculation [12]. The RI derived by this method was regarded as the primary reference when the D-P-LAVE(-) gave a wider RI as an indication of influence from latent abnormal values.

2.3.1.4. Parametric calculation using Tukey's outlier detection method: D-P-Tukey. This method assumes that there are no abnormal values within the Tukey range (LL ~ UL) specified by the formula below and thus regards the range as the RI [9]. The distribution of the source data was first made close to Gaussian by applying the Box-Cox transformation [X = $(x^{\lambda} - 1)/\lambda$] using the fixed power (λ) for each analyte as described above**.** By use of transformed values, 25 and 75 percentile points (Q_1^T, Q_3^T, respectively) was calculated. Subsequently, the following Tukey's outlier range ($LL^T \sim UL^T$) were regarded as corresponding to limits of RI in the transformed scale.

$$LL^T = Q_1^T - (Q_3^T - Q_1^T) \times 1.5$$

$$UL^T = Q_3^T + (Q_3^T - Q_1^T) \times 1.5$$

Finally, L^T and U^T were reverse transformed to obtain the limits of RI (L ~ U) in the original scale using the following formulae.

$$LL = (\lambda \times LL^T + 1)^{1/\lambda}$$

$$UL = (\lambda \times UL^T + 1)^{1/\lambda}$$

2.3.2. Indirect methods

2.3.2.1. Hoffman method: ID-Hoff. The assumption and procedures are the same as just described above except that quantile–quantile (Q-Q) plot was used and the linear segment was determined by visual inspection as in the original method [4]. Data was analyzed after the same transformation that was used for the Bhattacharya method (see below) with the same limitations on reporting. The analysis was performed using a spreadsheet application (Microsoft Excel) developed by one of the authors (G. Jones). In addition to the selection of λ, the selection of the range of data to include in the Gaussian model was made by the author.

2.3.2.2. Bhattacharya method: ID-Bhat. This method assumed that the distribution of the LIS source data consists of at least one Gaussian distribution, with the predominant one representing healthy subjects. A graphical method is used to identify this central distribution [5]. Data was analyzed with no transformation if a good fit with the data was obtained. Otherwise a Box-Cox transformation was used with either the λ identified from direct studies as described above, or a λ selected for the best fit. For some analytes a different λ was used for males and females. If the λ for best fit was a negative number, the RI was not derived as this implies a healthy subgroup could not be separated from those results likely to be affected by disease or other factors. The analysis was performed using a spreadsheet application (Microsoft Excel) developed by one of the authors (G. Jones). In addition to the selection of λ, the selection of bin size and bin location of the histogram and the range of data to include were made manually by the author.

2.3.2.3. Arzideh method: ID-TML. The method assumes that the central part of the distribution of patients' test results represents the non-disease ("healthy") population, which can be modelled by the power normal (PN) distribution family, if appropriately truncated on both sides of the peak, using a Box-Cox transformation function. $X = \frac{x^{\lambda}-1}{\lambda}$ (λ: power, *X*: power-transformed value *x*) [8].

The parameters of the PN distribution (μ, σ, λ) are estimated using the maximum likelihood method. A goodness-of-fit statistic (modified Kolmogorov-Smirnov (KS) statistic consisting of two terms, a two-sided KS for the goodness of the truncation interval and a one-sided KS for the goodness out of the truncation interval) was used to find an optimized interval for truncating the central part of the distribution. Then, the estimated PN distribution (characterized by λ, mean (μ), and SD (σ)) is used to determine the central 95% interval of the PN distribution representing the non-pathological values [8].

2.3.2.4. Wosniok method: ID-TMC. This truncated minimum chi-square (TMC) method is basically the same as the above-described method that depends on dissecting the central segment by fitting a PN distribution [9]. The procedures for parameter fitting were improved as follows. After stratification of source data by sex and age, an initial estimate of the PN distribution parameters λ, mean (μ), and SD (σ) were obtained by use of Q-Q plots based on Hoffmann's method [4,14]. The estimates were then optimized by adjusting the truncation interval in reference to the goodness-of-fit using chi-square statistics computed for the truncated segments. Finally, RIs for each stratum were estimated by use of the final fitted parameters as described above [8].

The members and associated members of C-RIDL collaborated and took charge of determining RIs by use of the following methods: T. Streichert, two German indirect methods (TML and TMC); G. Jones, indirect derivation using Hoffman and Bhattacharya methods; Y. Ozarda, direct derivation using Tukey's outlier detection method; K. Ichihara, direct derivation using a combination of parametric/nonparametric method with/without the LAVE method [11].

2.4. Assessment of differences in RI limits

To assess the practical significance of between-method differences in RIs, a bias ratio (BR) specified below was calculated at LL, median (Me), and UL of the RI determined by a given method.

$$BR_{LL} = \frac{LL - LL_0}{SD_{RI}}, BR_{Me} = \frac{Me - Me_0}{SD_{RI}}, BR_{UL} = \frac{UL - UL_0}{SD_{RI}}; SD_{RI} = \frac{UL_0 - LL_0}{3.92}$$

where LL_0, Me_0, and UL_0 respectively represent LL, Me, and UL determined by D-P-LAVE(-) that was set as a 'reference'. Because the denominator of the BR (SD_{RI}) corresponds to between-individual SD, its threshold was set to 0.375 according to the conventional specification of allowable bias at the minimum level [15]. As an exception to this scheme, D-P-LAVE(+) was set to a 'reference' when the effect of the LAVE method was not negligible in adjusting the RI limits in terms of BR.

3. Results

3.1. Data size of the LIS dataset before and after data cleaning

As a pre-processing of the LIS source dataset, two step cleansing procedures as recommended by Jones [3] was applied. The total data

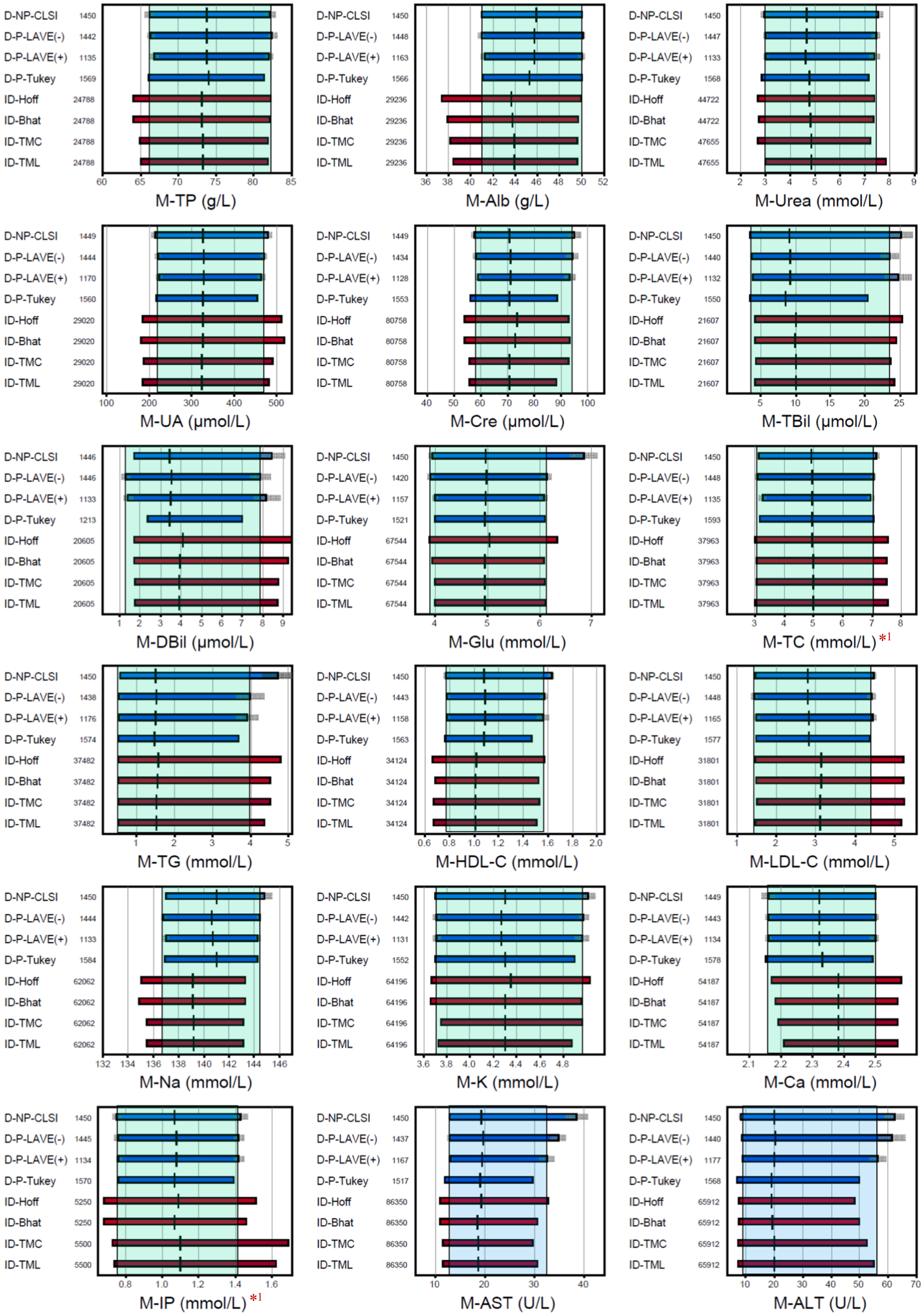
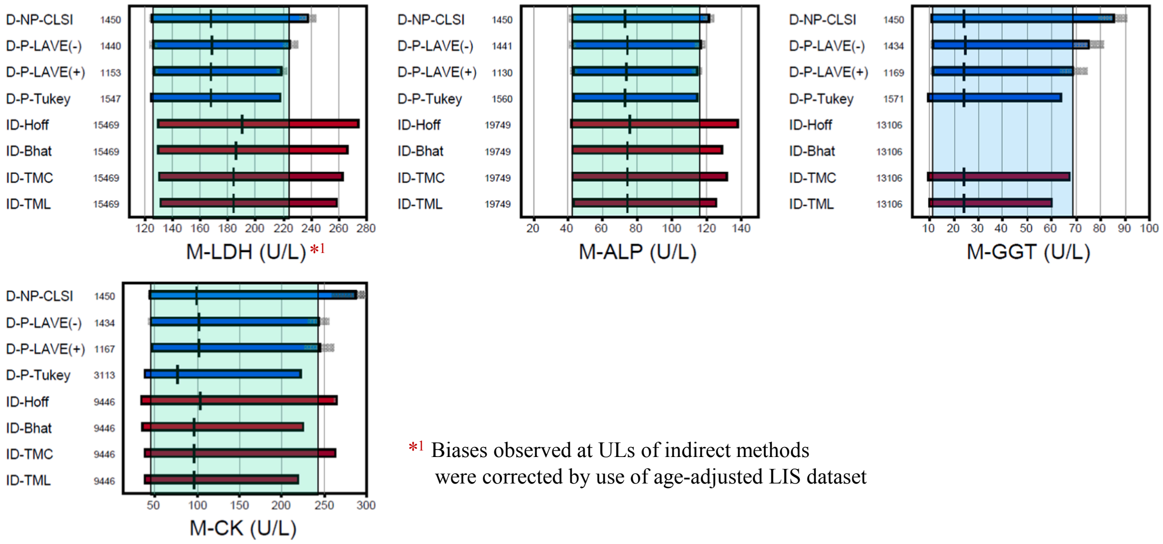


Fig. 1. Comparison of RIs for males by 8 calculation methods D-, ID- = direct, indirect methods. NP-, P- = nonparametric, parametric method; CLSI = CLSI guideline; LAVE = latent abnormal values exclusion method: Tukey = Tukey's outlier detection; Hoff = Hoffman method; Bhat = Bhattacharya method; TML = truncated maximum likelihood method by Arzideh; TMC = truncated minimum chi-square method by Wosniok; RIs derived for males (M) by 8 calculation methods were compared. RI by D-P-LAVE(-) is shown by green-shade as a reference for comparison for all analytes except for AST, ALT, and GGT, for which RI by D-P-LAVE (+) was set as a reference shown by a blue shade. RIs derived for females and males + females are shown collectively in **Suppl Fig. 2**. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



*1 Biases observed at ULs of indirect methods were corrected by use of age-adjusted LIS dataset

D-, ID- = direct, indirect methods. NP-, P- = nonparametric, parametric method;
CLSI = CLSI guideline; LAVE= latent abnormal values exclusion method: Tukey=Tukey outlier exclusion;
Hoff = Hoffman method; Bhat = Bhattacharya method; TMC=truncated minimum chi-square method;
TML = truncated maximum likelihood;

RIs derived for males (M) by 8 calculation methods were compared. RI by D-P-LAVE(-) is shown by green-shade as a reference for comparison for all analytes except for AST, ALT, and GGT, for which RI by D-P-LAVE(+) was set as a reference shown by a blue shade. RIs derived for females and males+females are shown collectively in Suppl Fig. 2.

Fig. 1. (*continued*).

size retrieved from the LIS for the 6-year period were 22,409,761. After limiting the data to those from outpatients and limiting special clinical departments as described in the Methods were 13,459,966 (60.1% of total data). To avoid duplicate records, the dataset was further restricted to allow only one record per year per patient. Hence, the final data size was 2,552,304 (11.4% of total data). The proportion of test results excluded per analyte by the cleaning procedure ranged from Mg (3.8%) to TC (33.8%). The stepwise changes in the data sizes are as shown in Table 2**.**

3.2. *Time-serial analysis for assessing between-year variations of the assays*

For confirming the long-term stability of assays, the LIS data for the study period of six-years (2011–2016) were retrieved after restricting to records that contained simultaneously tested results for ≥12 analytes out of the 25 targeted ones in this study. Then, each record was checked for a number of abnormal results in reference to the RIs currently in use in the laboratory. Any row of record with more than 3 abnormal results among the 25 analytes were deleted from the dataset. Then, this “multivariately normal” (MN) dataset was stratified by the date of measurement into every-six-month blocks and average of MN (AoMN) of each block was calculated for each analyte as tabulated in **Suppl. Table 3**. A bias of average of AoMN for the first two years (P1: 2011–2012), when the direct study was conducted, was computed from average of AoMN for the entire period (P2: 2011–2016) by the following formula, in analogy with the bias ratio (BR) described above:

$$\text{bias ratio (BR)} = \frac{|\text{average of AoMN for P1} - \text{average of AoMN for P2}|}{\text{SD}_{\text{RI}}}$$

where SD_{RI} represents standard deviation comprising the RI, or (UL − LL)/3.92.

A threshold for |BR| was again set to 0.375 based on the conventional specification of allowable analytical bias [15]. The fluctuations of AoMN were graphically analyzed as shown in **Suppl. Fig. 1**. Based on this analysis, we decided to exclude datasets for Cl, Mg, and AMY, with respective BR of 0.374, 0.865, and 0.386, from the subsequent analyses.

3.3. *Comparison of RIs across eight methods of calculation*

Suppl. Fig. 2 presents bar-chart comparison of the RIs derived by 8 methods for all 22 analytes in three ways for male (M) + female (F), M, and F. The representative graphs comparing RIs for males were shown in Fig. 1. The findings from this figure were presented as follow in three ways: comparison of RIs among direct methods, among indirect methods, and between direct and indirect methods. The evaluation of biases from RIs by D-P-LAVE(-) was made in reference to BRs at LL and/or UL shown in **Suppl. Tables 4**.

3.3.1. *Comparison of RIs among direct methods*

3.3.1.1. Nonparametric vs. Parametric method. RIs by NP and P methods matched well (|BRs| at LL or UL < 0.375) for analytes with low-skewness distribution (λ > 0.5): TP, Alb, urea, UA, Cre, TC, HDL-C, LDL-C, Na, K, Ca, and IP. Whereas, for analytes with skewed distributions (λ ≤ 0.5), ULs by NP method were shifted to the higher side, especially for Glu, TBil, TG, AST, ALT, LDH, GGT, and CK ($|\text{BR}_{\text{UL}}|$>0.5).

3.3.1.2. Effect of LAVE procedure. Effects of the LAVE method with lowering of ULs were observed for AST, ALT, and GGT, and slightly for UA, TG, and ALP in females, for LDH in males. This UL lowering effect of the LAVE procedure was consistent with those revealed in the interim report of the global study [13].

3.3.1.3. Performance of Tukey outlier detection method. The Tukey method (D-P-Tukey) using the direct study dataset gave nearly identical RIs as those by D-P-LAVE(+). Exception was, when the study λ adopted from the global study was not matched to the actual λ calculated by D-P-LAVE (+or -) method: i.e., the UL of D-P-Tukey was lower than D-P-LAVE(+) for Cre, DBil, and CK. This observation points to the importance of λ in approximating the Tukey's outlier range as the RI.

3.3.2. Comparison among indirect methods

RIs by the four indirect methods agreed quite well for TP, TC, HDL-C, and Na. For other analytes were generally inconsistent with each other in a variable degree. Whereas a variable degree of discrepancies were observed at LLs for TP, Alb, Cre, and Ca, and at ULs for the most other analytes.

Regarding Hoffmann and Bhattacharya methods, no results were obtained for GGT, and for creatinine (M + F) because of difficulty in identifying Gaussian peak by the manual procedure.

Comparing TML and TMC methods, the two methods gave very similar RIs except for urea, Cre, GGT, and CK due to a difference in algorithm of identifying the Gaussian peak. It is of note that in some analytes, the two methods estimated quite different λ values compared to those reported by the global study that are shown in **Suppl. Table 2**. This discrepancy was caused by dependence of power transformation on the width of excised peak segment (see Section 4).

3.3.3. Cross comparison between direct and indirect methods

By setting the RIs derived by the direct method: D-P-LAVE (-) [or D-P-LAVE(+) for AST, ALT, and GGT] for use as a primary reference, based on BR shown in **Suppl Table 4**, RIs by the indirect methods were compatible with those by the direct methods in the following 5 analytes: urea (M), TBil (M), Glu (M), K, and AST (F) with gender specificity indicated in the parenthesis. However, for the rest of analytes, a variety of shift patterns of RIs from the directly derived ones were observed. The patterns almost consistent across the indirect methods were as follows: (1) lower-side shift of LLs: Alb, Cre, HDL-C, and Na; (2) higher-side shift of ULs: DBil, TC, TG, LDL-C, ALT (F), LDH, and ALP; (3) both-sided shift of RI limits with extended width of the RI: UA and IP (M). As a whole, the proportions of biases observed either at LL or UL by each method was tabulated at the bottom of **Suppl Table 4**. The overall proportion of biases for the four indirect methods was 72% (ID-Hoff 81%; ID-Bhat 76%; ID-TMC 68%; ID-TML 62%).

3.3.4. Effect of adjusting LIS dataset for age and study period on RIs

As shown in Table 1, there were some biases in age distribution of the LIS dataset compared to the dataset obtained in the direct study: average age for males and females were respectively 39.6 and 38.8 years for the direct study dataset, and 45.4 and 47.5 years for the LIS dataset. Besides, year-to-year bias of minor degree may exist even after exclusion of three analytes (Cl, Mg, and AMY) that showed appreciable between-year variations in test results. Therefore, we prepared an age-adjusted LIS dataset for each analyte by counting data sizes of every 5-year age-strata, and randomly deleting data from a stratum with excessive data size compared to the direct study dataset. In the process of adjustment, the period of data was also limited to 2011–2012.

RIs were recalculated using the age-adjusted dataset by the TML method (ID-TMLadj), which was regarded as a representative indirect method. The difference in BR (ΔBR) between ID-TML and ID-TMLadj was evaluated for all analytes as shown in **Suppl. Table 5**. By regarding $|\Delta BR| \geq 0.375$ as a significant change, BR-UL was lowered below the threshold of 0.375 (○) for TC, IP, and LDH in males, and for UA, ALT, LDH, ALP and CK in females. Overall, the proportion of biases after excluding those analytes was decreased appreciably from 72 to 47% (ID-Hoff 51%, ID-Bhat 54%, ID-TMC 44%, and ID-TML 38%). This finding points to the importance of prior matching of age-distribution of LIS data, which tend to have a higher-side age bias, for improved validation of RIs by the indirect method (see Section 4).

3.4. Distributions of sources data used in direct and indirect methods

To explain the biases of RIs by indirect methods, distributions of datasets used in the indirect and direct methods were graphically compared using two graph modes (Fig. 2). One drawn by the line chart represents pre-cleaned LIS dataset served for indirect methods, and the other drawn by the solid-bar histogram represents the dataset obtained from the direct study. The red vertical line indicates the median of the direct-study dataset, which helps identify the bias of a presumed center of "non-disease" group of the indirect dataset.

Four patterns were distinguished from the spread and peak of the line-chart histogram relative to the bar histogram: (1) a lower-side spread with peak shift (observed for TP, Alb, HDL-C, and Na). (2) a higher-side spread with/without shift in peak (LDL-C, Ca, LDH, and GGT / TBil, DBil, and TG), (3) a both-side spread without peak shift (UA, TC, IP, ALP, and CK), and (4) matched shape of the two histograms (urea, Cre, Glu, K, AST, and ALT).

These four shapes of the indirect dataset explain some of the biases observed in RIs predicted by the indirect methods: i.e., the pattern-1 led to lowered LLs of Alb, HDL-C, and Na; the pattern-2 led to raised ULs of DBil, TG, LDL-C, Ca, and LDH; the pattern-3 led to lowered LLs plus raised ULs of UA and IP (M), and also to raised ULs of TC, ALP, and CK.

Whereas, for the pattern-4, RIs by the indirect methods generally agreed well for urea, Glu, and K.

4. Discussion

Requirements for a scheme of validating indirectly derived RIs is (1) to conduct a direct and indirect study in parallel with measurements done in the same clinical laboratory, and (2) to make "non-disease" individuals to be data-mined from the LIS comparable to the healthy subjects to be recruited.

We had an opportunity of conducting such a validation study according to the scheme, although we did the indirect study retroactively. For the direct part of the study, we carried it out in accordance with the up-to-date harmonized protocol, elaborated by C-RIDL [11], with recruitment of 3066 healthy volunteers nationwide and centralized/standardized measurements for 25 chemistry tests [10]. No regional differences were observed in any analytes. For the indirect study, we tried to make use of the LIS source data of six years (2011–2016) that covered the period of the direct study (2011–2012). The comparability of test results over the 6-year period was confirmed, after exclusion of three analytes that showed significant between-year bias, as shown in **Suppl. Table 3**. The range of age for the LIS data was restricted to 18–65 years, but an average age was higher by 5.8 years in males and 8.7 years in females compared to the direct study, in which age distribution was deliberately made nearly flat. However, we did not adjust the LIS dataset having assumed that the influence of the age bias in determining the RIs would be small. Nonetheless, after completion of determining RIs by all methods, we retroactively examined an effect of adjusting age-distribution of the LIS dataset by use of TML method, regarded as representing the indirect method. As a result, we recognized that the age-adjusted dataset led to correction of the biases observed in ULs for TC and LDH in both sexes, for UA, ALT, ALP, and CK in females, and for IP in males as shown in **Suppl. Table 5**. Overall frequency of RIs with biases by four indirect methods was reduced from 72% to 47%. This appreciable drop clearly indicates that the age-adjustment should have been done for stringent assessment of the indirect method. However, it may not be practical to prepare fully age-adjusted dataset as a prerequisite for applying the indirect method.

In this study, we regard it imperative for the indirect method to transform the central part of distributions of LIS dataset as close to Gaussian as possible. In Hoffmann and Bhattacharya methods, the Box-Cox power transformation method was used with its power (λ) adopted with some modification from the average λ values reported in the IFCC global study [11]. Whereas optimal λ values were predicted as an

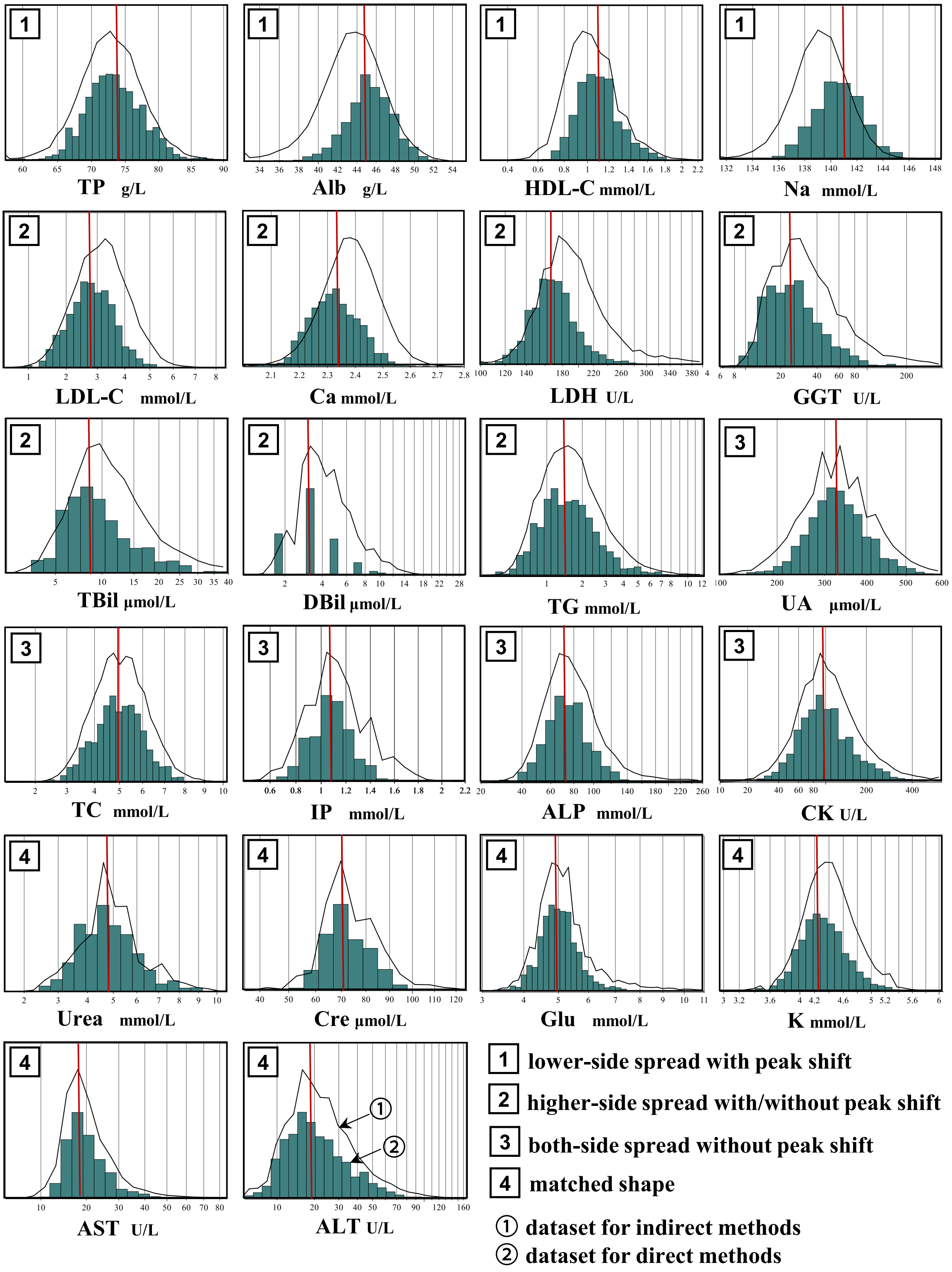


Fig. 2. Comparison of datasets used in direct and indirect methods. Male datasets used for derivation of RIs by direct and indirect methods were compared by histograms: ① for dataset used by indirect methods, ② used by direct methods. The height of ② was exaggerated for ease of comparing the shapes. The red line represents the median of the direct dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

integral part of the TML and TMC methods.

It is of noteworthy that TML and TMC sometimes gave a low λ value for analytes with near Gaussian distributions (Na, Ca) or slightly skewed distributions (LDH, ALP). A common feature of the analytes is that their distributions start up far away from the origin (0.0) with displaced peak. Because TML and TMC methods used the one-parameter Box-Cox formula without the origin of transformation, the predicted λ value for those special analytes can vary widely retaining comparable degree of fitness to the Gaussian distribution: i.e., both λ = 0.0 and λ = 1.0 give nearly identical RIs for those peak-displaced distributions. Therefore, λ values by the two indirect methods are valid, but not relevant in interpreting the skewness of the distribution when the peak is displaced. On the other hand, the direct method using the two-parameter Box-Cox formula, because of optimization of the transformation origin, λ varies in accordance with the skewness of the distribution.

After exclusion of RIs that were influenced by the age-bias of the LIS dataset, systematic biases remain observed in 47% of RIs by the indirect methods. The biases can be categorized into the following three patterns in reference to the altered histogram peak profiles of indirect datasets as shown in Fig. 2.

The first pattern of the bias was lower-side shift of predicted LL observed for Alb and Na. It is attributed to high frequency of low values in the LIS dataset, which occur at proximity to the presumed non-disease group and obviously lowered the location of the distribution peak.

The second pattern was a higher-side shift of the predicted UL observed for TG and LDL-C. Again, high frequency of abnormal values near ULs displaced the distribution peak upwards. The third pattern was outward shift of both LL and UL. It implies that abnormal results occur fairly frequently on both sides of the non-disease group in close proximity, the situation of which caused a broader distribution peak of the LIS dataset. This bias pattern of RIs was observed in UA and IP before correction of the age-bias, but were not found among RIs after age-adjustment.

As an exception to the above influence of altered distribution peaks in determining RIs, prediction biases were also observed for Cre, AST (M), and ALT despite matched histogram profiles of indirect dataset to the direct one (the fourth pattern in Fig. 2). The predicted RIs for these analytes by the four indirect methods were grossly inconsistent among each other with RI limits either lower or higher than those of the direct method RIs. The common features of these analytes are (1) skewed distributions with long tailing to the higher side, and (2) matched peaks of the two dataset. From Fig. 2, it is notable that overlapping of abnormal clusters with non-disease population is less intense. Therefore, dissecting the "non-disease" group seems to be relatively easy. However, its efficacy depends on the algorithm and the appropriateness of the power (λ), which explain the inconsistency among the four methods.

The propensity of the prediction bias of RIs by the indirect methods, despite prior two-step cleaning of the source data as recommended [3], clearly points to the need for more rigorous attempt to clean up the LIS source data. Besides, in this validation study, we just evaluated indirect methods that all relied on truncation of values in univariate distributions. For more efficient data cleaning, it should be necessary to refer to other relevant information multivariately in the data selection process.

In the literature, there have been several attempts to filter out values of healthy subjects from LIS source data. In 1996, Ichihara and Kawai [16] reported derivation RIs for 13 plasma proteins by use of health screening dataset, which turned out to contain sizable number of abnormal results. Therefore, they adopted multivariate scheme of excluding any individuals with abnormal results in 25 clinical chemistry screening tests that had been measured simultaneously, based on correlations among plasma proteins and screening tests. It was a forerunner or non-iterative version of LAVE method. In 1998, Ritchie et al. reported determination of RIs for serum proteins by use of LIS dataset that contained diagnostic classification of each patient in 90 categories. Selection of appropriate categories led to improved RIs [17]. In 2005 REALAB study [18], RIs for 23 basic tests were derived from LIS dataset after excluding any record with abnormal results in related tests, and also after restricting any record from an individual who was tested more than once over three years period. The study group found the latter exclusion scheme was effective in restricting results from non-healthy individuals.

In 2015, Yamakado et al. [19] derived RIs for 18 standardized chemistry tests and 8 hematology tests from 1.5 million health screening database accumulated nationwide in Japan. They first excluded 80% of individuals on the basis of either BMI $\geq$ 25 kg/m^2, smoking habit, ethanol >20 g/day, regular medication, or high blood pressure, and then, approximately 50% of the remaining individuals were further excluded by applying the LAVE method most strictly without allowing any abnormal value in related tests. Thus, only 10% of "supernormal" individuals remained for derivation of RIs. The resultant RIs were found quite comparable to the Japanese common RIs determined by a nationwide study of direct approach [20].

In the presence of these well controlled studies of indirect approaches, it will be necessary to conduct an additional study to investigate whether or not more rigorous attempt to clean the LIS source data by multifaceted approach can eliminate systematic biases observed among the indirect methods.

5. Limitations

This study was launched with a primary objective of validating a group of indirect methods that have been recommended in the recent review paper on the indirect method [3]. Therefore, we just covered the four indirect methods that all rely on univariate delineation of presumed "non-disease" group, ignoring the presence of other reports that relied on multifaceted information in performing data mining. In fact, we are currently in a process to develop such an indirect method featuring sex and age matched retrieval of LIS data and multivariate based scheme of data cleaning. Therefore, we regard this study as a preliminary step toward that goal.

6. Conclusion

We investigated the validity of RIs determined by four major indirect methods that rely univariately on the shape of central distribution peak of LIS source dataset. In order to make the RIs established by the nationwide "direct" study for use as a reference, LIS source data were retrieved from the same clinical laboratory for the matched period. RI predicted by four indirect methods showed a variable degree of biases at either LL or UL or both. By comparison of histograms of the direct and indirect datasets, it was revealed that the biases were associated with altered shape and location of the distribution peaks of LIS dataset: i.e., lower-side peak shift led to lowered LL (eg. Na, Alb, etc), higher-side peak shift led to raised UL (eg. TG, LDL-C, etc), and broadened peak led to both-side biases (eg. UA, IP). However, the biases observed for TC, ALP, etc. were attributed to inevitable age-bias of the LIS dataset. After recalculation using age-adjusted dataset, the overall proportion of biased RIs was decreased from 72% to 47% (38–54%). To reduce the biases in predicting the RIs, it is necessary to apply more rigorous attempt of data-cleaning in multivariate manner in reference to other relevant clinical/laboratory information.

CRediT authorship contribution statement

Yesim Ozarda: Writing - original draft, Writing - review & editing, Methodology, Validation, Supervision. **Kiyoshi Ichihara:** Writing - original draft, Methodology, Writing - review & editing, Validation. **Graham Jones:** Formal analysis, Writing - review & editing. **Thomas Streichert:** Formal analysis, Writing - review & editing. **Robab Ahmadian:** Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are very thankful, for the success of this study, to the earnest contributions of local coordinator: Dr. Melahat Dirican (Uludag University School of Medicine, Central Laboratory, Bursa, Turkey).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2021.05.030.

References

[1] IFCC and CLSI. EP28-A3c document, Defining, establishing and verifying reference intervals in the clinical laboratory: approved guideline, third ed., vol. 28 No. 30, 2010.

[2] Y. Ozarda, K. Ichihara, J.H. Barth, G. Klee, on behalf of the CRIDL on Reference intervals and Decision Limits, IFCC. Protocol and standard operating procedures for common use in worldwide multicenter study on reference values, Clin. Chem. Lab. Med. 51 (2013) 1027–1040.

[3] G.R. Jones, R. Haeckel, T.P. Loh, K. Sikaris, T. Streichert, A. Katayev, J.H. Barth, Y. Ozarda, Indirect methods for reference interval determination – review and recommendations, Clin. Chem. Lab. Med. 57 (2019) 20–29.

[4] R.G. Hoffmann, Statistics in the practice of medicine, JAMA 185 (1963) 864–873.

[5] C.G. Bhattacharya, A simple method of resolution of a distribution into gaussian components, Biometrics 23 (1967) 115–135.

[6] A. Katayev, J.K. Fleming, D. Luo, et al., Reference intervals data mining: no longer a probability paper method, Am. J. Clin. Pathol. 143 (2015) 134–142.

[7] A. Katayev, C. Balciza, D.W. Seccombe, Establishing reference intervals for clinical laboratory test results: is there a better way? Am. J. Clin. Pathol. 133 (2010) 180–186.

[8] F. Arzideh, W. Wosniok, E. Gurr, W. Hinsch, et al., A plea for intra-laboratory reference limits. Part 2. A bimodal retrospective concept for determining reference limits from intra-laboratory databases demonstrated by catalytic activity concentrations of enzymes, Clin. Chem. Lab. Med. 45 (2007) 1043–1057.

[9] W. Wosniok, R. Haeckel, A new indirect estimation of reference intervals: truncated minimum chi-square (TMC) approach. Clin. Chem. Lab. Med. 2019 Jul 4. pii: /j/cclm.ahead-of-print/cclm-2018-1341/cclm-2018-1341.xml. doi: http://dx.doi.10.1515/cclm-2018-1341.

[10] Y. Ozarda, K. Ichihara, D. Aslan, N. Ilhan, D. Sadak-Atali, et al., A multicenter nationwide reference intervals study for common biochemical analytes in Turkey using Abbott analyzers, Clin. Chem. Lab. Med. 52 (2014) 1823–1833.

[11] 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 467 (2017) 70–82.

[12] 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 (2010) 1537–1551.

[13] G. Klee, K. Ichihara, Y. Ozarda, N.A. Baumann, J. Straseski, et al., Reference intervals: comparison of calculation methods and evaluation of procedures for merging reference measurements from two US medical centers, Am. J. Clin. Pathol. 150 (2018) 545–554.

[14] G. Hoffmann, R. Lichtinghagen, W. Wosniok, Simple estimation of reference intervals from routine laboratory tests, J. Lab. Med. (2016), https://doi.org/10.1515/labmed-2015-0104.

[15] C.G. Fraser, Biological Variation: From Principles to Practice, AACC Press, Washington, DC, 2001.

[16] 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 (1996) 110–117.

[17] R.F. Ritchie, G.E. Palomaki, L.M. Neveux, et al., Reference distributions for immunoglobulins A, G, and M: a practical, simple, and clinically relevant approach in a large cohort, J. Clin. Lab. Analy. 12 (1998) 363–370.

[18] E. Grossi, R. Colombo, S. Cavuto, C. Franzini, The REALAB project: a new method for the formulation of reference intervals based on current data, Clin. Chem. 51 (2005) 1232–1240.

[19] M. Yamakado, K. Ichihara, Y. Matsumoto, Y. Ishikawa, et al., Derivation of gender and age-specific reference intervals from fully normal Japanese individuals and the implications for health screening, Clin. Chim. Acta 447 (2015) 105–114.

[20] K. Ichihara, Y. Yamamoto, T. Hotta, S. Hosogaya, et al., Collaborative derivation of reference intervals for major clinical laboratory tests in Japan, Ann. Clin. Biochem 53 (2016) 347–356.