Universiteit Leiden
The Netherlands

Harmonization of indirect reference intervals calculation by the Bhattacharya method

Martinez-Sanchez, L.; Gabriel-Medina, P.; Villena-Ortiz, Y.; Garcia-Fernandez, A.E.; Blanco-Grau, A.; Cobbaert, C.M.; ... ; Elzen, W.P.J. den

Citation
Martinez-Sanchez, L., Gabriel-Medina, P., Villena-Ortiz, Y., Garcia-Fernandez, A. E., Blanco-Grau, A., Cobbaert, C. M., ... Elzen, W. P. J. den. (2022). Harmonization of indirect reference intervals calculation by the Bhattacharya method. *Clinical Chemistry And Laboratory Medicine*, *61*(2), 266-274. doi:10.1515/cclm-2022-0439

Version:	Publisher's Version
License:	Licensed under Article 25fa Copyright Act/Law (Amendment Taverne)
Downloaded from:	https://hdl.handle.net/1887/3512263

Note: To cite this publication please use the final published version (if applicable).

Luisa Martinez-Sanchez*, Pablo Gabriel-Medina, Yolanda Villena-Ortiz, Alba E. García-Fernández, Albert Blanco-Grau, Christa M. Cobbaert, Daniel Bravo-Nieto, Sarai Garriga-Edo, Clara Sanz-Gea, Gonzalo Gonzalez-Silva, Joan López-Hellín, Roser Ferrer-Costa, Ernesto Casis, Francisco Rodríguez-Frías and Wendy P.J. den Elzen

Harmonization of indirect reference intervals calculation by the Bhattacharya method

https://doi.org/10.1515/cclm-2022-0439
Received May 5, 2022; accepted November 3, 2022;
published online November 17, 2022

Abstract

Objectives: The aim of this study was to harmonize the criteria for the Bhattacharya indirect method Microsoft Excel Spreadsheet for reference intervals calculation to reduce between-user variability and use these criteria to calculate and evaluate reference intervals for eight analytes in two different years.

Methods: Anonymized laboratory test results from outpatients were extracted from January 1st 2018 to December 31st 2019. To assure data quality, we examined the monthly results from an external quality control program. Reference intervals were determined by the Bhattacharya method with the St Vincent's hospital Spreadsheet firstly using original criteria and then using additional harmonized criteria defined in this study. Consensus reference intervals using the additional harmonized criteria were calculated as the mean of four users' lower and upper reference interval results. To further test the operation criteria and robustness of the obtained reference intervals, an external user validated the Spreadsheet procedure.

Results: The extracted test results for all selected laboratory tests fulfilled the quality criteria and were included in the present study. Differences between users in calculated reference intervals were frequent when using the Spreadsheet. Therefore, additional criteria for the Spreadsheet were proposed and applied by independent users, such as: to set central bin as the mean of all the data, bin size as small as possible, at least three consecutive bins and a high proportion of bins within the curve.

Conclusions: The proposed criteria contributed to the harmonization of reference interval calculation between users of the Bhattacharya indirect method Spreadsheet.

Keywords: Bhattacharya; harmonization; indirect approach; reference intervals.

Introduction

Reference intervals are very important, as they support clinical decision making based on laboratory results [1, 2].

Luisa Martinez-Sanchez, Pablo Gabriel-Medina and Yolanda Villena-Ortiz contributed equally to this work as first authors.

Fracisco Rodríguez–Frías and Wendy den Elzen contributed equally to this work as senior authors.

***Corresponding author: Luisa Martinez-Sanchez**, Biochemistry Department, Clinical Laboratories, Vall d'Hebron University Hospital, Barcelona, Spain; Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Spain; Department of Clinical Chemistry and Laboratory Medicine, Leiden University Medical Centre, Leiden, The Netherlands; and Clinical Biochemistry Research Team, Vall d'Hebron Institute of Research (VHIR), Barcelona, Spain, Phone: +34 649445158,
E-mail: luisa.maria.martinez.lm@gmail.com. https://orcid.org/0000-0002-3936-0156

Pablo Gabriel-Medina, Yolanda Villena-Ortiz, Albert Blanco-Grau and Francisco Rodríguez-Frías, Biochemistry Department, Clinical Laboratories, Vall d'Hebron University Hospital, Barcelona, Spain; Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Spain; and Clinical Biochemistry Research Team, Vall d'Hebron Institute of Research (VHIR), Barcelona, Spain. https://orcid.org/0000-0003-0925-0498 (Y. Villena-Ortiz)

Alba E. García-Fernández, Daniel Bravo-Nieto, Sarai Garriga-Edo, Clara Sanz-Gea, Gonzalo Gonzalez-Silva, Joan López-Hellín and Roser Ferrer-Costa, Biochemistry Department, Clinical Laboratories, Vall d'Hebron University Hospital, Barcelona, Spain; and Clinical Biochemistry Research Team, Vall d'Hebron Institute of Research (VHIR), Barcelona, Spain

Christa M. Cobbaert, Department of Clinical Chemistry and Laboratory Medicine, Leiden University Medical Centre, Leiden, The Netherlands

Ernesto Casis, Clinical Laboratories, Vall d'Hebron University Hospital, Barcelona, Spain

Wendy P.J. den Elzen, Clinical Biochemistry Research Team, Vall d'Hebron Institute of Research (VHIR), Barcelona, Spain; and Department of Clinical Chemistry, Amsterdam UMC, University of Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

Laboratory test results outside the reference interval could be defined as pathological and may warrant further attention [2]. In addition, the accuracy of in-range values is also important, since the unjustified absence of medical actions could also drive negative long-term consequences for patients [3]. Therefore, establishing correct, updated and specific reference intervals for our population is a critical point in the clinical laboratory and assuring their quality and reliability is one of the most important tasks of specialists in clinical laboratory medicine.

Reference intervals are currently calculated using the direct approach [4]. The limitations and disadvantages of this methodology have been widely discussed [5]. To note: Complexity to select, contact and enrol 120 healthy random individuals, especially for those tests that require multiple partitions per sex and age or the costs of performing the study; among others. These drawbacks may become so tedious that routine laboratories frequently choose to adopt the reference intervals suggested by the manufacturer, calculated using a different population and settings.

Given these limitations, indirect methods have emerged as an alternative approach [6–10] and are increasingly used. These methods use data from thousands of individuals from already performed routine analyses, collecting the data from the laboratory information system (LIS) and subsequently analysing them statistically.

Availability of a high number of test results in the LIS is an essential requirement for the calculation of reference intervals by indirect approaches. Clinical Laboratory Vall d'Hebron is one of the largest laboratories in Europe by workload and complexity as more than 60,000 tests results are produced every day and the catalogue includes more than 1,000 tests, providing "*in vitro* diagnostics" service to the majority of the Barcelona city public health activity. Faced with this scenario, we consider the calculation of reference intervals by indirect approaches a positive and revolutionary opportunity in our laboratory.

In 2019, the International Federation of Clinical Chemistry (IFCC) published a review encouraging clinical laboratories to participate in the development of indirect methods for reference intervals determination [11]. Multiple methods have been developed using the idea of calculating reference intervals from patient populations: Hoffmann [12], Pryce [13], Bhattacharya [14], NUMBER [8], kosmic [15], truncated minimum chi-squared (TMC) [16], among others. In this study the Bhattacharya method was used due to a free access tool available online that facilitate the handling of indirect methods (often highly complex statistically) for non-statistical experts in the laboratory.

The Bhattacharya method was described in 1967 [14] and is a graphical method for identifying a Gaussian distribution (reference population) in the midst of a complete dataset with both reference individuals and non-reference individuals (non-healthy subjects). Two requirements are necessary to separate these two populations mathematically: 1) they do not highly overlap and 2) the total sample size is large enough (more than 1,500 in the original description). In the original description of the method, the Gaussian distribution of data was considered another requirement. Since most laboratory data do not show a normal distribution, Baadenhuijsen et al. and Oosterhuis et al. described some modifications to address some of these limitations [17, 18].

The currently used Spreadsheet and other online applications for the Bhattacharya method apply linear regression to shape a line of best fit for the segment that the user visually chooses as a straight-line. This line identifies the reference population. Actually, more robust and reliable reference intervals are estimated if larger numbers of individuals are included (more than 5,000) and a greater proportion of the dataset is from the reference population [11].

In the present study, Bhattacharya analyses were performed using St Vincent's hospital Spreadsheet available online (http://www.sydpath.stvincents.com.au/). This method requires the (subjective) input of the user for selecting an appropriate bin size and the points included on the graph [19]. The first purpose of this study was to study between user variability when calculating reference intervals in the Excel application; then a second purpose was to standardize the criteria initially defined to reduce between-user variability in the reference interval results. Then, a third aim was to calculate and evaluate reference intervals for eight tests during two different years (2018 and 2019), based on the new criteria.

Materials and methods

Data selection

Anonymized laboratory test results from individuals (more than 18 years old) visiting general practitioners were extracted from January 1st 2018 to December 31st 2019 from the LIS of the Clinical Laboratory Vall d'Hebron in Barcelona.

Test results from outpatients belonging to primary care attention centres were included, since we expected a high proportion of healthy people. Haemolytic (>0.03 mmol/L haemoglobin), lipaemic (>0.45 mmol/L Intralipid®) and icteric (>23.94 µmol/L bilirubin) samples were excluded. A total of 1,067,794 clinical requests were selected (509,408 from 2018 and 558,386 from 2019). A detailed description of the dataset from 2018 is provided elsewhere [20].

Analytical measurements

Samples were collected from 62 blood collection centres and were transported via eight different routes to the laboratory (average transportation time 3 h). Serum tubes included separating gel and coagulation activator (BD Vacutainer®). The samples were transported to the laboratory in cool boxes with a temperature monitoring system. After arriving in the laboratory, the samples were centrifuged either 12 min at 3,500 rpm (2,438*g*) when handled manually outside the track or 10 min at 3,000 rpm (2,113*g*) when on the track. No clinically significant differences in the test results were found when comparing the two centrifugation conditions (results not shown).

Biochemistry tests were measured on AU5800 chemistry analysers (Beckman Coulter®). The following test methods were used according to the instructions for use of the manufacturer: alanine aminotransferase (ALT), IFCC recommended method without pyridoxal phosphate traceable to Beckmann coulter master calibrator; glucose, reaction with hexokinase traceable to NIST SRM 965; calcium, reaction with arsenazo III traceable to NIST SRM 909bL1; magnesium, direct method with xylidyl blue traceable to NIST SRM 909bL2; inorganic phosphorus, reaction with ammonium molybdate traceable to Beckmann coulter master calibrator; chloride, potassium and sodium by indirect ion selective electrodes traceable to NIST SRM 919, 918 and 919 respectively.

Quality assessment

To assure data quality, we examined the monthly results from the biochemistry specific external quality control program from the Spanish Society of Laboratory Medicine (Sociedad Española de Medicina de Laboratorio, SEQC^{ML}). In this scheme, the results from the external quality control materials obtained in our laboratory were compared with the average calculated from every laboratory participating in the program using the same analytical method and/or instrument. Alike routine laboratory practice, when our result was within two times the standard deviation from other laboratories participating in the scheme using the same method, data from this particular month and test were accepted as valid. If our result exceeded ± two standard deviations, we excluded the data from that particular test, month and instrument. In addition, uncertainty was calculated as the sum of standard uncertainty from the calibrator material, the analytical coefficient of variation and the uncertainty of the analytical system.

To assess longitudinal accuracy across lot numbers, daily averages of the extracted General Practitioner test results were investigated to check for analytical stability over time. Averages were calculated per batch of 200 results a day and were visually compared with the average per month and average for the whole year 2018 or 2019.

Reference intervals calculation and statistical analysis

The Bhattacharya method was performed to determine the reference intervals using the programmed Microsoft Excel sheet by St Vincent's hospital (available in: http://www.sydpath.stvincents.com.au/) as advised by the IFCC Committee on Reference Intervals and Decision Limits (C-RIDL) [11]. A workflow with the main steps followed during the project process is presented in Figure 1.

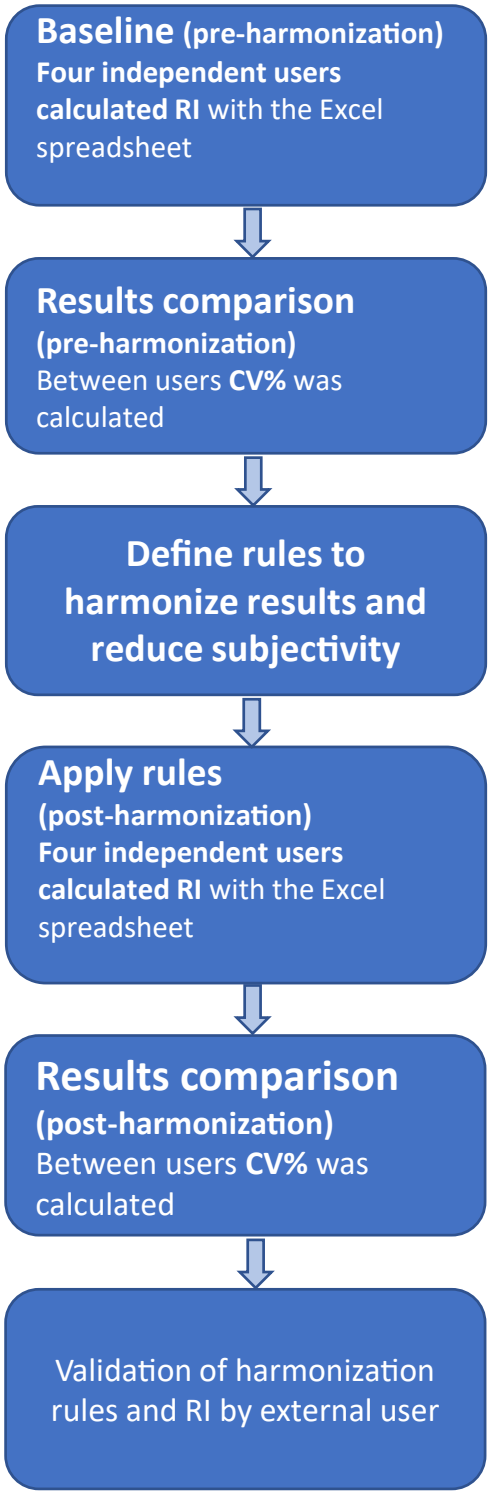
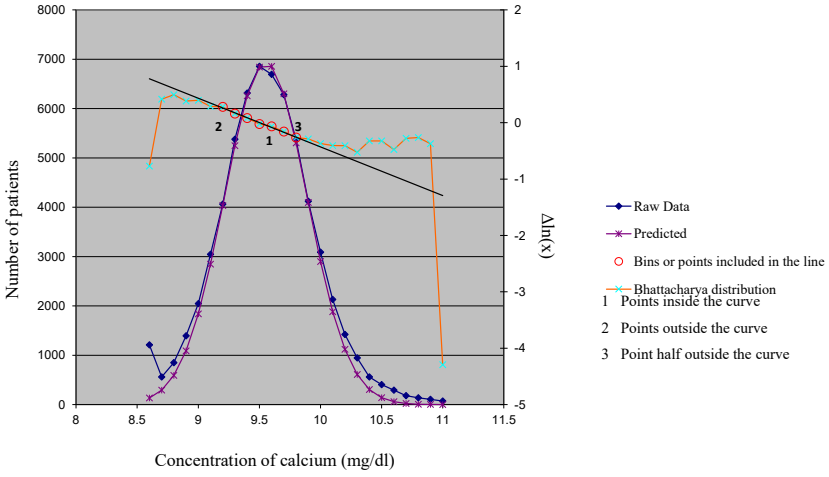


Figure 1: Workflow with the main steps followed during the project.

First, identical excel templates were made, avoiding errors in data transfer. Then, four different laboratory specialists (users) worked independently to obtain the reference intervals following the initial recommendations from the original sheet. The variability between users was compared and evaluated using the coefficient of



Δln(x): Difference between ln(x) values of the cumulative difference of patients

Figure 2: Representation of the graphs obtained using the St Vincent's hospital excel Spreadsheet (http://www.sydpath.stvincents.com.au/) for Bhattacharya indirect method reference interval calculation.
Raw data present how the original data is distributed according to the selected bin size; predicted data present how the data would be distributed according to the bin size selected and to the number of bins included in the line and considered for the calculations.

variation (CV) of the reference interval limits calculated by magnitude and partition. To simplify results presentation, the CV is shown together for 2018 and 2019 and separating low and high interval per test. For ALT only the high interval is shown as the low intervals were not considered clinically relevant.

To reduce variability between users, we focused on the user-dependent variables based on the results obtained, i.e. in the bin size, the total number of bins included in the reference population (#points) and the number of bins graphically inside the curve (which was considered 0.5 when half bin stood outside the curve) (Figure 2). Based on our experiences with the Spreadsheet, we developed additional new consensus criteria for the use and operation of the excel sheet in order to reduce inter-user variability. Again, the same four users obtained the reference intervals independently using the new consensus criteria.

We calculated the mean between the four users and the 95% confidence interval (CI) using the formula $\mu \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, being μ=mean, $Z_{\alpha/2}$=1.96 and σ=standard deviation (SD) with n=4 for the four independently obtained low and high limits of the reference intervals. If, after this calculation, any of the four users, either the low or high limit of the calculated reference interval lay outside the 95% CI, then its results were considered not valid and discarded for the final calculation of the new 95% CI and reference intervals. Consensus reference intervals were calculated as the mean of the valid users' high and low results. As an example, the results obtained for the high limit for potassium in mmol/L in 2018 were: user 1=4.99; user 2=5.03; user 3=5.11 and user 4=5.03. The 95% CI calculated for the four users (n=4) was 4.99–5.04. Therefore, as the result from user three was higher than the 95% CI, it was considered not valid and the mean reference interval was re-calculated using the remaining three users' results.

Per test and per group boxplots were visually inspected to decide whether or not subgroup differentiated reference intervals were necessary per sex and age.

Results evaluation

To test for significant differences between pre and post harmonization strategies, the F-test [21] was applied to the SD before and after harmonization. In addition, to decide upon the acceptance of the obtained CV after harmonization within-individual biological variation was used [22].

To reduce variability of the user dependent variables, statistical correlations between them were analysed. After applying the new criteria, the number of data (n), the number of decimal points and the central bin (defined by the data) were considered as independent variables. The contribution of the independent variables on the bin size and the bins included in the line (#points) were analysed in univariate and multivariate models using linear regression analysis and Pearson correlation.

To further test the operation criteria and robustness of the obtained reference intervals, an external user reproduced the Spreadsheet procedure applying the defined criteria.

Flagging rates per test and per year were calculated with an independent dataset from primary care (1st January 2020 to 31st December 2020). Results are shown as the percentage of individuals outside the reference intervals.

Results

The data obtained fulfilled the quality criteria and, therefore, were included in the present study. Longitudinal accuracy was also considered fulfilled as observed in annual averages for the eight laboratory tests studied (Supplementary Table 1) and in the monthly averages plots. An example of these plots is shown in Figure 3 for potassium, where increase in potassium concentrations was observed during colder temperatures months [23].

Table 1 shows the coefficient of variation (CV) between users' reference intervals calculations when applying the original criteria (pre-harmonization), median (Q1–Q3) of 3.99% (1.49–10.95). The additional criteria defined to standardize the analysis by reducing between user variability in the reference intervals calculation and their justifications are shown in Table 2. The CVs after applying these additional

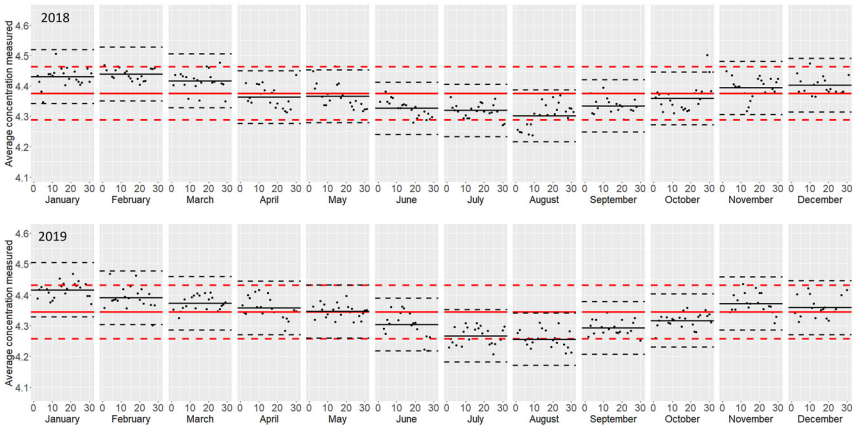


Figure 3: Daily averages of consecutive primary attention patient results in the extracted datasets (black dots) for potassium results in 2018 and 2019. Monthly average (black) and annual average (red) are represented as lines and biological variation percentage over and under the mean is represented a slashed lines.

Table 1: Coefficient of variation (CV) between the reference intervals results calculated between the four users per each test in 2018 and 2019 for the upper reference limit (URL) and the lower reference limit (LRL) with the original criteria of Microsoft Excel Bhattacharya Spreadsheet (pre-harmonization) and with the criteria proposed within this study (post-harmonization). Within-subject biological variation (CVi) and p-value for the F-test between the pre and post harmonization results are also shown.

Analytical test	RV 2018 and 2019	CV_i, %	CV, % pre-harmonization	CV, % post-harmonization	p-Value (F-test)
Sodium	LRL	0.5	0.19	0.21	0.809
	URL		0.35	0.34	0.931
Potassium	LRL	4.0	0.71	0.74	0.923
	URL		1.16	0.78	0.297
Chloride	LRL	1.0	0.77	0.00	**<0.001**
	URL		0.85	0.00	**<0.001**
Calcium	LRL	1.8	1.30	0.00	**<0.001**
	URL		1.11	0.00	**<0.001**
Magnesium	LRL	2.8	0.73	0.00	**<0.001**
	URL		0.94	0.00	**<0.001**
Phosphate (males 18–50 years)	LRL	7.7	4.00	4.18	0.869
	URL		2.23	2.12	0.877
Phosphate (females 18–50 years)	LRL		3.39	4.73	0.402
	URL		1.79	1.77	1.000
Phosphate (males 51–65 years)	LRL		5.61	0.00	**<0.001**
	URL		2.23	0.00	**<0.001**
Phosphate (females 51–65 years)	LRL		1.22	0.00	**<0.001**
	URL		0.93	0.00	**<0.001**
Phosphate (males >65 years)	LRL		5.86	4.41	0.490
	URL		1.90	2.06	0.844
Phosphate (females >65 years)	LRL		3.25	3.82	0.734
	URL		2.32	2.26	1.000
Glucose	LRL	4.9	4.29	3.36	0.679
	URL		3.42	0.75	**0.001**
ALT (males 18–50 years)	URL	10.0	15.23	9.44	0.123
ALT (females 18–50 years)	URL		13.64	5.71	**0.018**
ALT (males 51–65 years)	URL		7.32	7.71	0.719
ALT (females 51–65 years)	URL		11.04	1.04	**<0.001**
ALT (males 66–80 years)	URL		5.99	7.65	0.711
ALT (females 66–80 years)	URL		5.79	3.45	0.126
ALT (males >80 years)	URL		6.35	8.38	0.603
ALT (females >80 years)	URL		5.09	9.02	0.331
Median (Q2)			**2.28**	**1.41**	
Lower quartile (Q1)			**1.07**	**0.00**	
Upper quartile (Q3)			**5.66**	**4.24**	

Significant p-values and summary results are highlighted in bold.

Table 2: Summary of original and additional criteria defined to harmonize the analysis by reducing between user variability in the reference intervals calculated using St Vincent's hospital Spreadsheet available online (http://www.sydpath.stvincents.com.au/).

Original criteria	Additional criteria	Explanation
Initial central bin (including log transformed data) should be close to the mean or median.	Set the value of the central bin as the mean of all the data.	Central bin could be fixed as the arithmetic mean if there is a little influence from pathological results in the database and this would reduce the variability.
Bin size must be equal to or larger than the reporting interval.	To adjust the bin size, use the value of the reporting interval of the data as a starting point and increase it to meet all the following criteria. Finally select the smallest possible bin size.	Higher bin sizes lead to low resolution graphs and inappropriate reference interval results.
Select data from four to six bins to include in the Bhattacharya analysis.	The line must be defined with a minimum of 4 bins, at least three of them consecutively.	The biggest possible number of bins should be selected, since it allows a larger population to be included in the calculation of reference intervals.
	If bins are not considered for the adjustment of the line, they must be placed between two included bins.	Excluding intermediate bins assumes that the subpopulation is not homogeneous with respect to the bins immediately nearby. Excluding a single bin might be permissible, if a minimum bin size is selected. Excluding more than one intermediate bin would be an error and would skew the result.
The Bhat line must be very straight. Particularly data points "steeper" than the line of best fit should be included.	R-squared value 0.99 is big enough.	A larger R-squared does not modify or ensure validity of the results obtained. Instead, looking for a larger R-squared can penalize the selection of the most important variables, the bin size and the number of bins selected.
	The maximum number of points on the line must be included within the curve.	Points included inside the curve highlight the importance of the central bunch of data for the final calculation of reference intervals, in contrast with the data found in the extremes of the distribution.
If in doubt, seek expert advice and/or another operator for validation.	The Spreadsheet should be explored by independent scientists (four in our case).	Reducing the inherent subjectivity that could lead to less reliable results when obtained by a single user.

criteria are also presented in Table 1 (post-harmonization). Considerably less inter-user variability was obtained in the post-harmonization results, with a median (Q1–Q3) of 2.90% (0.06–7.35). Table 1 also shows within-subject biological variation for comparison with the obtained CV and the p-value from the F-test to assess significant differences between variation of the pre and post harmonization results.

Results of the bin size, the number of total points included in the line (#points) and the number of points within the distribution curve (#points inside) obtained by each of the four users are shown in Supplementary Table 2.

Supplementary Table 3 shows the reference intervals results obtained in the years 2018 and 2019 per user, the final 95% CI between the users and the reference intervals currently used in our laboratory derived from analyser inserts (RI_{cu}) for the eight tests studied. Shaded results in Supplementary Table 3 were considered not valid (outside 95% CI, n=4) and discarded for the final calculation of the final 95% CI (n=3) and reference intervals.

A linear correlation was found by the univariate model between the bins or points included in the line (see Figure 2, y=#points) and the number of available data (x=n) ($r=0.277$; $p<0.001$). This correlation was defined by the formula: $y=5.818 + 0.53 \times 10^{-5}$ x. According to that, when calculating reference intervals for a laboratory test with for example 10,000 results, the recommended points or bins over the curve based on our formula are 5.8, rounded to six included bins. It means that there is a proportional increase in #points with higher n. In the multiple linear regression analysis, bin size was statistically associated with the central bin ($\beta=0.071$, $p<0.001$) and decimal points (-0.0943, $p<0.001$) ($R^2_{adjusted}=0.706$). It is important to remark that the observed correlations are specific for the selected analytes, the units and the methodology.

The external user results differed in nine out of the 80 limits calculated by the initial users (Supplementary Table 3). This was particularly the case for the lower and upper limits for chloride of 98 mmol/L and 110 mmol/L

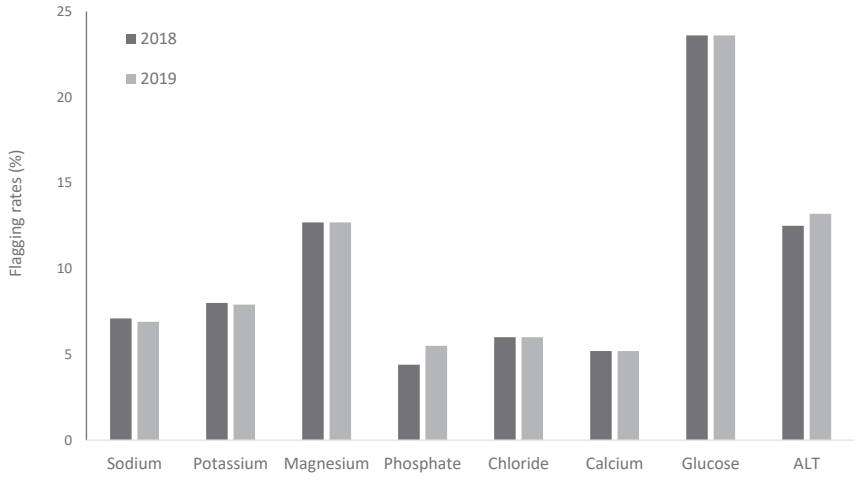


Figure 4: Validation of the calculated RIs in the 2020 laboratory dataset. Percentage of flagged patient results outside the calculated reference intervals in 2018 and 2019.

respectively, the lower limit for magnesium of 0.78 mmol/L (1.9 mg/dL), the lower limit for phosphate in males (>65 years) 0.68 mmol/L (2.1 mg/dL) from 2018, the upper limit of phosphate in males (>65 years) of 1.42 mmol/L (4.4 mg/dL) from 2019, the lower limit for ALT in males (51–65 years) 9 U/L for 2018, the higher limit for ALT in males (>80) 29 U/L for 2018 and the lower limits for ALT in females (18–50) 5 U/L and (51–65) 7 U/L from 2018. The remaining reference intervals calculated by the external user fell within the 95% CI calculated by the independent users.

Results of the flagging rates obtained by applying the reference intervals calculated for 2018 and 2019 in a population sample from 2020 are shown in Figure 4. The percentages of the flagging rates for 2018 and 2019 exceeded 5% of pathological values for all analytes except for phosphate in 2018.

Discussion

The Bhattacharya indirect method for reference intervals calculation can be performed in a simple and easy way using the Spreadsheet created by St Vincent's hospital. Initial recommendations, included in the "instruction" sheet from the excel Spreadsheet, allows the user to obtain reliable results but variability between users was found to be an issue (Table 1). Additional criteria and recommendations created within this study from the observation of these variations (Table 2), reduced subjectivity when performing the procedure for reference intervals calculation using the Spreadsheet. In addition, we observed a dependency between some (subjective) decisions the user has to face (as the number of points to be included in the line or the bin size) and other known variables as the number of results or the test units. When these relationships are taken into account, even less between user variability may be observed.

A reduction in within user variation using the Excel spreadsheet were gained in 22 out of 32 reference interval limits presented in Table 1; 13 of those reductions were statistically significant. For the 10 cases in which a reduction was not observed, the change in CV was not statistically significant. A reduction of CV appeared challenging for ALT for which extreme values are found often and a non-Gaussian distribution is present for the test results. As the Bhattacharya method does not exclude extreme values previous to reference intervals calculation, medical tests with a high proportion of extreme values will have more variability between users when using the tool. Between user variations for the calculated reference intervals were always lower than the within-subject biological variation for both pre and post harmonization, except for ALT (males and females, 18–50 and females 51–65) where CV pre harmonization was higher than 10%. To note, the pre harmonization CVs were in general very close to the within-individual biological variation threshold for those tests in which we gained a significant reduction of CV for the post harmonization results by applying the additional harmonized criteria.

In previous studies other procedures to exclude bin selection, based on the differences in data frequency between consecutive bins, have been proposed; either by establishing a minimum data frequency regarding the mode or by graphical observation of the residues obtained from Bhattacharya graphic against test concentration [24]. Since we aimed to propose a simple and objective method, this was not considered in this study.

Bin size is also an important variable for method performance. Smaller bin sizes will lead to higher random

variation in the number of data per bin and therefore the complexity of the linear fit that represents the Gaussian population will be higher [25]. We noticed that the reporting interval of data is also an important source of variability between users (Supplementary Table 2). A lower reporting interval of data leads to different bin sizes between users and therefore more variability in obtained reference intervals. Potassium is an example of this.

The ratio of results outside the calculated reference intervals (flagging rates) in 2018 and 2019, from a new dataset with population from 2020 has shown similar results for all analytical tests. Therefore, even with slight numeric differences in reference intervals, the method leads to coherent results attending to the ratio of pathological population detection. The expected results higher than 5% in flagging rates (based on the statistical model of reference intervals where 95% of healthy population are within the intervals [26]) are accomplished in all cases except for phosphate in 2018 (4.4%).

It is important to remark that the same dataset from 2018 was used in a previous study [20] for calculating reference intervals using two indirect methods: The Dutch NUMBER method [11] and the German reference limit estimator method [16]. The calculated reference intervals were comparable with the mentioned results [20] for all the included analytical test. Comparison with other important reference interval studies such as CALIPER (direct method) [27], ARIA (indirect method) [28] and NORIP (direct method) [29] also gives comparable results for all tests, except for ALT. This is an important topic of further research.

One of the weaknesses in our study is that level one commutable external quality control was not applied yet in our laboratory in 2018 and 2019. Two important aspects from data quality should be always considered for data reuse: 1) the use of methods traceable to higher order reference materials and 2) the use of level one commutable external quality control. The fulfilment of these two requirements is a prerequisite for the application of calculated reference intervals to the clinical practice [5]. If data quality is assured [30], the obtained reference intervals from different populations can be universally compared. The proposed methodology for the use of the Spreadsheet in Bhattacharya calculation is useful for data from several laboratories where these conditions are met.

To conclude, we assessed between user variability when using the Bhattacharya Excel Spreadsheet and designed additional criteria to harmonize reference intervals calculation. Considering the eight laboratory tests analysed, we conclude that the proposed additional criteria for the use of St Vincent's hospital Spreadsheet contribute to the harmonization of reference intervals calculation by the Bhattacharya method. This system, including the additional criteria presented, could be applied in other clinical laboratories to optimize reference intervals calculation by the Bhattacharya method.

Acknowledgments: We thank Dr. Raymond Noordam for the R script to design the figures for the moving averages analysis.
Research funding: None declared.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Not applicable.
Ethical approval: Not applicable.

References

1. Siest G, Henny J, Gräsbeck R, Wilding P, Petitclerc C, Queraltó JM, et al. The theory of reference values: an unfinished symphony. Clin Chem Lab Med 2013;51:47–64.
2. Sikaris KA. Physiology and its importance for reference intervals. Clin Biochem Rev 2014;35:3–14.
3. Placzkowska S, Terpinska M, Piwowar A. The importance of establishing reference intervals – is it still a current problem for laboratory and doctors? Clin Lab 2020;1:1429–38.
4. Committee S, Section C, Petitclerc C, Biochimie SD, Montreal HD. International Federation of Clinical Chemistry (IFCC) 1), 2) approved recommendation (1987); on the theory of reference values part 2. Selection of individuals for the production of reference values International Federation of Clinical Chemistry. J Clin Chem Clin Biochem 1987;25:2–30.
5. Martinez-Sanchez L, Marques-Garcia F, Ozarda Y, Blanco A, Brouwer N, Canalias F, et al. Big data and reference intervals: rationale, current practices, harmonization and standardization prerequisites and future perspectives of indirect determination of reference intervals using routine data. Adv Lab Med/Av en Med Lab 2020;0:9–16.
6. Poole S, Schroeder LF, Shah N. An unsupervised learning method to identify reference intervals from a clinical database. J Biomed Inform Elsevier Inc 2016;59:276–84.
7. Lykkeboe S, Nielsen CG, Christensen PA. Indirect method for validating transference of reference intervals. Clin Chem Lab Med 2018;56:463–70.
8. Den Elzen WPJ, Brouwer N, Thelen MH, Le Cessie S, Haagen IA, Cobbaert CM. NUMBER: standardized reference intervals in The Netherlands using a "big data" approach. Clin Chem Lab Med 2019;57:42–56.
9. Clerico A, Trenti T, Aloe R, Dittadi R, Rizzardi S, Migliardi M, et al. A multicenter study for the evaluation of the reference interval for TSH in Italy (ELAS TSH Italian Study). Clin Chem Lab Med 2019;57:259–67.
10. Shaw JLV, Cohen A, Konforte D, Binesh-Marvasti T, Colantonio DA, Adeli K. Validity of establishing pediatric reference intervals

based on hospital patient data: a comparison of the modified Hoffmann approach to CALIPER reference intervals obtained in healthy children. Clin Biochem The Canadian Society of Clinical Chemists 2014;47:166–72.
11. Jones GRD, Haeckel R, Loh TP, Sikaris K, Streichert T, Katayev A, et al. Indirect methods for reference interval determination – review and recommendations. Clin Chem Lab Med 2019;57: 20–9.
12. Hoffmann RG. Statistics in the practice of medicine. JAMA, J Am Med Assoc 1963;185:864–73.
13. Pryce JD, Lond MD. Level of haemoglobin in whole blood and red blood-cells, and proposed convention for defining normality. Lancet 1960;2:333–6.
14. Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. Biometrics 1967;23:115.
15. Zierk J, Arzideh F, Kapsner LA, Prokosch HU, Metzler M, Rauh M. Reference interval estimation from mixed distributions using truncation points and the Kolmogorov-Smirnov distance (kosmic). Sci Rep 2020;10:1704.
16. Wosniok W, Haeckel R. A new indirect estimation of reference intervals: truncated minimum chi-square (TMC) approach. Clin Chem Lab Med 2019;57:1933–47.
17. Baadenhuijsen H, Smit JC. Indirect estimation of clinical chemical reference intervals from total hospital patient data: application of a modified Bhattacharya procedure. Clin Chem Lab Med 1985;23: 829–40.
18. Oosterhuis WP, Modderman TA, Pronk C. Reference values: Bhattacharya or the method proposed by the IFCC? Ann Clin Biochem 1990;27:359–65.
19. Sikaris KA. Separating disease and health for indirect reference intervals. J Lab Med 2021;45:55–68.
20. Martinez-Sanchez L, Cobbaert CM, Noordam R, Brouwer N, Blanco-Grau A, Villena-Ortiz Y, et al. Indirect determination of biochemistry reference intervals using outpatient data. PLoS One 2022;17:e0268522.
21. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. Crit Rev Clin Lab Sci 1989;27:409–37.
22. Aarsand AK, Fernandez-Calle P, Webster C, Coskun A, Gonzales-Lao E, Diaz-Garzon J, et al. The European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). [cited 2022 Sep 25]. Available from: https://biologicalvariation.eu/.
23. Davis KR, Crook MA. Seasonal factitious increase in serum potassium: still a problem and should be recognised. Clin Biochem 2014;47:283–6.
24. Hemel JB, Hindriks FR, Van Der Slik W. Critical discussion on a method for derivation of reference limits in clinical chemistry from a patient population. J Automat Chem 1985;7:20–30.
25. Farrell CL, Nguyen L. Indirect reference intervals: harnessing the power of stored laboratory data. Clin Biochem Rev 2019;40:99–111.
26. Wayne P. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline, 3rd ed. CLSI document EP28-A3c; Wayne, PA: Clinical and Laboratory Standards Institute; 2008. C28–A3.
27. Adeli K, Higgins V, Nieuwesteeg M, Raizman JE, Chen Y, Wong SL, et al. Biochemical marker reference values across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian Health Measures Survey. Clin Chem 2015;61:1049–62.
28. Tate JR, Sikaris KA, Jones GR, Yen T, Koerbin G, Ryan J, et al. Harmonising adult and paediatric reference intervals in Australia and New Zealand: an evidence-based approach for establishing a first panel of chemistry analytes. Clin Biochem Rev 2014;35:213–35.
29. Rustad P, Felding P, Franzson L, Kairisto V, Lahti A, Mårtensson A, et al. The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties. Scand J Clin Lab Invest 2004;64:271–84.
30. Jansen RTP, Cobbaert CM, Weykamp C, Thelen M. The quest for equivalence of test results: the pilgrimage of the Dutch Calibration 2.000 program for metrological traceability. Clin Chem Lab Med 2018;56:1673–84.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2022-0439).