Andrea Padoan*, Aldo Clerico, Martina Zaninotto, Tommaso Trenti, Renato Tozzoli, Rosalia Aloe, Antonio Alfano, Sara Rizzardi, Ruggero Dittadi, Marco Migliardi, Marcello Bagnasco and Mario Plebani

Percentile transformation and recalibration functions allow harmonization of thyroid-stimulating hormone (TSH) immunoassay results

https://doi.org/10.1515/cclm-2019-1167
Received November 11, 2019; accepted December 9, 2019

Abstract

Background: The comparability of thyroid-stimulating hormone (TSH) results cannot be easily obtained using SI-traceable reference measurement procedures (RPMs) or reference materials, whilst harmonization is more feasible. The aim of this study was to identify and validate a new approach for the harmonization of TSH results.
Methods: Percentile normalization was applied to 125,419 TSH results, obtained from seven laboratories using three immunoassays (Access 3rd IS Thyrotropin, Beckman Coulter Diagnostics; Architect System, Abbott Diagnostics and Elecsys, Roche Diagnostics). Recalibration equations (RCAL) were derived by robust regressions using bootstrapped distribution. Two datasets, the first of 119 EQAs, the second of 610, 638 and 639 results from Access, Architect and Elecsys TSH results, respectively, were used to validate RCAL. A dataset of 142,821 TSH values was used to derive reference intervals (RIs) after applying RCAL.
Results: Access, Abbott and Elecsys TSH distributions were significantly different ($p < 0.001$). RCAL intercepts and slopes were −0.003 and 0.984 for Access, 0.032 and 1.041 for Architect, −0.031 and 1.003 for Elecsys, respectively. Validation using EQAs showed that before and after RCAL, the coefficients of variation (CVs) or among-assay results decreased from 10.72% to 8.16%. The second validation dataset was used to test RCALs. The median of between-assay differences ranged from −0.0053 to 0.1955 mIU/L of TSH. Elecsys recalibrated to Access (and vice-versa) showed non-significant difference. TSH RI after RCAL resulted in 0.37–5.11 mIU/L overall, 0.49–4.96 mIU/L for females and 0.40–4.92 mIU/L for males. A significant difference across age classes was identified.
Conclusions: Percentile normalization and robust regression are valuable tools for deriving RCALs and harmonizing TSH values.

Keywords: laboratory information systems; percentile normalization; recalibration equation; reference intervals; statistical harmonization; thyroid disease; thyroid-stimulating hormone; thyrotropin.

Introduction

According to all international guidelines, measurement of serum thyroid-stimulating hormone (TSH) is considered the first-line screening test for thyroid dysfunction

***Corresponding author: Andrea Padoan,** Department of Medicine (DIMED), University of Padova, via Giustiniani 2, 35128, Padova, Italy; and Department of Laboratory Medicine, University-Hospital of Padova, via Giustiniani 2, 35128, Padova, Italy, E-mail: andrea.padoan@unipd.it. https://orcid.org/0000-0003-1284-7885
Aldo Clerico: Laboratory of Cardiovascular Endocrinology and Cell Biology, Department of Laboratory Medicine, Fondazione CNR-Regione Toscana Gabriele Monasterio, Scuola Superiore Sant'Anna, Pisa, Italy
Martina Zaninotto: Department of Laboratory Medicine, University-Hospital of Padova, Padova, Italy
Tommaso Trenti: Dipartimento di Medicina di Laboratorio e Anatomia Patologica, Azienda Ospedaliera Universitaria e USL di Modena, Modena, Italy
Renato Tozzoli: Clinical Pathology Laboratory, Department of Laboratory Medicine, Azienda per l'Assistenza Sanitaria n.5, Pordenone Hospital, Pordenone, Italy
Rosalia Aloe: Dipartimento di Biochimica ad Elevata Automazione, Dipartimento Diagnostico, Azienda Ospedaliero-Universitaria di Parma, Parma, Italy
Antonio Alfano: Clinical Pathology, Hospital ASL TO4, Ciriè, Turin, Italy
Sara Rizzardi: Laboratorio Analisi Aziendale (SC), Azienda Socio-Sanitaria Territoriale di Cremona, Istituti Ospitalieri, Cremona, Italy
Ruggero Dittadi: U.O.C. Laboratorio Analisi, Ospedale dell'Angelo, AULSS3 Serenissima, Mestre, Venezia, Italy
Marco Migliardi: S.C. Laboratorio Analisi, A.O. Ordine Mauriziano di Torino, Turin, Italy
Marcello Bagnasco: Università degli Studi di Genova, Genova, Italy
Mario Plebani: Department of Medicine (DIMED), University of Padova, Padova, Italy; and Department of Laboratory Medicine, University-Hospital of Padova, Padova, Italy. https://orcid.org/0000-0002-0270-1711

Published online January 13, 2020

throughout all the lifespan (including pregnancy, postpartum and neonatal periods), for the evaluation of thyroid hormone replacement in patients with primary hypothyroidism, and for the assessment of suppressive therapy in patients with follicular cell-derived thyroid cancer [1–3]. Currently, levels of TSH are routinely assayed with non-competitive immunoassay methods [1]. Although the analytical performance of the TSH immunoassays has progressively improved in the last 30 years [1], there are still some systematic differences between the commercially available methods [4, 5]. Systematic bias or difference in interferences between TSH immunoassays may produce misleading interpretation when samples of the same patients are measured by different methods, especially in patients treated with drugs which are able to affect thyroid function or to cause interferences in immunoassay systems [6–9].

In 2010 [10–12] and 2014 [13], the IFCC Working Group for Standardization of Thyroid Function Tests published several studies concerning the standardization of both TSH and thyroid hormone immunoassay methods. More recently, the same group reported the results of another study on the evaluation and harmonization, rather than standardization, of TSH immunoassay methods, commercially available at that moment [14]. The aim of the study was to promote harmonization of immunoassay methods for TSH, because a process of standardization was not considered possible owing to the lack of accepted reference measurement procedures (RPMs) for this hormone [14]. Indeed, consensus has been progressively accumulated in the last few years about the importance of a more global approach to achieve a better harmonization in laboratory medicine, especially in the field of the most popular immunoassay methods [15–22].

A multicenter study, performed on behalf of the Italian Section of the European Ligand Assay Society (ELAS), evaluated the systematic differences between the most popular TSH immunoassays in Italy [5]. The conclusions of this study were that it is possible to produce a mathematical approach, which can obtain a better harmonization between the different TSH methods [5]. After recalibration, performed using a mathematical approach based on the principal component analysis, the variation of TSH values significantly decreased from a median pre-calibration value of 13.53% (10.79%–16.53%) to 9.63% (6.90%–13.21%).

Another important clinical problem related to the routine use of TSH assays is the large between-method difference in reference interval (RI) values [1]. The upper limits of the reference population of TSH immunoassays are strongly affected by outlier values, related to individuals with thyroid autoimmunity (thyroid autoantibody positive) or sub-clinical thyroid disease. Other factors related to population demographics (such as age, sex, and ethnicity), iodine intake, BMI, smoking status and administration of some drugs can affect the serum TSH levels and generate false-positive TSH elevations as well [1, 6, 7]. Further, analytical interferences such as heterophile antibodies have been reported to affect the reproducibility of results from healthy population [23]. For these reasons, the accurate evaluation of TSH reference values requires the enrollment of a very large number of rigorously screened normal euthyroid volunteers [1, 24]. Considering these inclusion and exclusion criteria for the enrollment of the reference population, as well as the systemic bias between TSH immunoassay methods, it is not surprising that RIs for TSH have remained poorly defined [2, 13].

For the calculation of RIs for TSH, some recent studies have suggested some experimental approaches based on indirect RI calculation, using very large populations with the aim of reducing these drawbacks [25–29]. However, robust statistical analyses, including resampling approaches, are needed for the accurate exclusion of all possible outliers [30, 31]. Theoretically, data mining approaches may be less accurate in the calculation of TSH reference values than the experimental studies based on large reference population. As a result, reference intervals, calculated with data-mining techniques, usually require an independent and accurate evaluation of their clinical effectiveness and efficiency using specific clinical studies.

In 2017, the Italian section of the European Ligand Assay Society (ELAS) organized a multi-center study (named ELAS TSH Italian Study) among several Italian clinical laboratories for the evaluation of TSH RIs using large laboratory databases. Preliminary results obtained from four Italian clinical laboratories, using the same method for the measurement of serum TSH confirmed that data-mining techniques can be used to calculate clinically useful RIs for TSH [29]. Prompted by these preliminary results, the authors extended the experimental protocol of the TSH Italian Study to include data related to three different TSH immunoassay methods collected by seven Italian clinical laboratories.

The aim of the present study was to develop and validate a new approach for harmonizing TSH results, based on regression equations and bootstrapped statistical methods. For this purpose, the TSH results from a large database of seven clinical laboratories were used [29]. Results were then validated on a different already

published database to check the performances of RCALs. Finally, age- and gender-specific RIs of TSH were derived after recalibration.

Materials and methods

Population database

The TSH values stored in the Laboratory Information System (LIS) of clinical laboratories of seven Italian city hospitals were analyzed. TSH measurements performed in samples collected from individuals referred by primary care practitioners throughout a period of about 2 years (2016–2017) were recorded by the LIS of the seven clinical laboratories. According to the clinical and demographic information available on the LIS, pregnant women were excluded from the study. The TSH results of individuals with free thyroxine (FT4), free triiodothyronine (FT3) and thyroid-autoantibodies outside the RIs of the methods used in the laboratories participating in the study were also excluded. Additionally, individuals with laboratory data (if available) suggesting the presence of thyroid or pituitary disease, a history of abnormal thyroid function test results or drug assumption, which can alter thyroid function test results, were excluded. Only one TSH value per individual was considered. Finally, individuals with laboratory test results suggesting acute or chronic diseases of cardiac, lungs, renal or liver systems were also excluded. These TSH data, listed in an Excel file with an alphanumeric barcode and together with the relative sex and age values of individuals, were sent to the reference laboratory of the study (i.e. Laboratory of the Fondazione CNR Regione Toscana G. Monasterio). The seven clinical laboratories used different alphanumeric bar codes in order to render unidentifiable the individual personal data to the investigators of the reference laboratory. These TSH measurements (with the respective age and sex-related values of individuals enrolled in the study) constituted the original database for the present ELAS TSH Italian Study.

According to this experimental protocol, this study was performed in accordance with the 1964 Helsinki declaration and its later amendments (the 2013 revision) or comparable ethical standards. All the clinical laboratories participating in the studies followed the recommendations of their Institutional Ethical Committees regarding the privacy-preserving data mining.

TSH assay

Three different immunoassay methods for TSH measurement were evaluated in the present study: Access TSH (3rd IS) Thyrotropin (REF B63284) by Beckman Coulter Diagnostics (distributed in Italy by Beckman Coulter Italia S.p.A, Cassina de' Pecchi, Milano, Italy); Architect System it TSH (REF 7K62) by Abbott Diagnostics (distributed in Italy by Abbott Diagnostics Italia SrL, Roma, Italy); Elecsys TSH (REF 07028091190) by Roche Diagnostics (distributed in Italy by Roche Diagnostics Italia S.p.A., Monza, Italy). The TSH measurements were performed in the laboratories according to the instructions suggested by the manufacturers.

Statistical analyses

Standard statistical analyses were carried out using the JMP program (version 12.1.0, SAS Institute Inc., SAS Campus Drive, Cary, NC, USA) and R (version 3.5.2 – "Eggshell Igloo", R Foundation for Statistical Computing). Base 10 logarithmic transformation ($\log_{10}$) of data were used for TSH. Considering the set of data, all values near the LoD value of TSH immunoassay methods (≤0.01 mIU/L) and those >20 mIU/L were excluded because these values are considered to be in the hyperthyroid and hypothyroid ranges. The remaining possible outliers were detected by the Tukey test using the formula: $cTnI > Q_3 + 3$ IQR, as the gating parameter, where Q_3 and IQR, respectively, are the third quartile and interquartile range ($Q_3 - Q_1$) of TSH distribution.

Percentile transformation was used to transform $\log_{10}$-transformed TSH data. Briefly, data were first transformed into percentiles. Then, the percentiles from different distributions were paired to obtain a table. This table was used to calculate robust regressions, using Huber M-estimation by the package MASS in R [32]. This process was iterated by bootstrapping. Briefly the bootstrapped distribution, obtained by resampling of 5000 observations for each immunoassay TSH distribution, was calculated and used as the merged distribution to derive RCALs. This process was iterated 1000 times and the results were averaged for obtaining the RCAL estimates. The obtained RCALs were used to recalibrate the TSH results of each immunoassay method in order to better harmonize the results. The exact k-sample permutation test was performed by the package Perm in R by using 10,000 Monte Carlo replications in order to reduce the influence of outliers and 95% confidence intervals of p-values. Percentiles comparisons were performed by the R function "pairwisePercentileTest" of the Rcompanion Package, by using R = 10,000 iterations. Dunn's test or Wilcoxon's test was used to evaluate the group's differences for non-Gaussian distributions [33]. Cubic smoothing spline was used to evaluate the association between the age and $\log_{10}$ TSH values.

Results

Distributions of TSH values

The descriptive statistics of the distributions of TSH values obtained by the LIS of the clinical laboratories of seven Italian city hospitals is reported in Table 1. After $\log_{10}$ transformation, the overall data showed a deviation from linearity in the upper and lower boundaries (Supplementary Figure 1). Exact permutation tests showed statistical significances between Access and Architect ($p = 0.0002$, 95% CI: 0.0001 and 0.0007), between Access and Architect ($p = 0.0002$, 95% CI: 0.0001 and 0.0010) and between Access and Elecsys ($p = 0.001$, 95% CI: 0.0001 and 0.0101). Furthermore, descriptive statistics were used for the bootstrapped distribution that was generated by random sampling with re-substitution of 5000 observations for each method, which were all combined together to obtain a final merged distribution of 15,000 TSH values. Table 1 also reports the claimed manufacturers' RIs for TSH values

Table 1: Descriptive statistics of TSH values (mIU/L) measured by the three immunoassay methods tested in this study and of the bootstrapped distribution.

Distribution of TSH values, mIU/L							Manufacturers' claimed RI
Data	Size, n	Mean	Median	SD	2.5th pct (90% CI)[a]	97.5th pct (90% CI)[a]	
Access	89,657	1.985	1.705	1.233	0.403 (0.399–0.408)	5.224 (5.166–5.272)	0.45–5.33[b]
Architect	14,109	1.756	1.528	1.055	0.406 (0.390–0.420)	4.512 (4.420–4.620)	0.35–4.94[c]
Elecsys	21,653	2.090	1.820	1.266	0.420 (0.410–0.430)	5.460 (5.380–5.560)	0.270–4.20[d]
Bootstrapped distribution	15,000	1.944	1.674	1.1967	0.411 (0.400–0.420)	5.114 (5.030–5.210)	–

Manufacturers' reference intervals (RIs) were also reported. SD, standard deviation; 95% CI, 95% confidence intervals; SE, standard error estimated by SD of the Monte Carlo Results. [a]Estimated by non-parametric bootstrap method, with M = 1000 replicates; [b]estimated by using approximately 400 subjects of a general population of approximately equal numbers of males and non-pregnant females between the ages of 21–88; [c]estimated by using 549 reference subjects with normal free T4; [d]estimated by 2.5th and 97.5th percentiles of 516 reference subjects.

(mIU/L) obtained by the following inserts: Access, version B83033 D August 2016; Architect, version G3-3871/R04 2012-09 and Elecsys version 07028091500-2.0 2017-09.

Estimation of recalibration equations

The equations for recalibrating each immunoassay method (RCAL) are reported in Table 2. For each method, these equations were obtained by using the percentile value of $\log_{10}$-transformed TSH values as the dependent variable and the percentile value of the resulting merged distribution as the independent variable. The slopes and intercepts were estimated by averaging the iterated bootstrapping results. Equations from Table 2 can be used to recalibrate the immunoassay in order to improve the harmonization of the results.

Validation of the obtained recalibration equations

In order to validate the previously identified recalibration regressions, data from 119 EQAs samples, distributed to more than 200 Italian clinical laboratories in 2012–2015 annual cycles of the Immunocheck study, were used. The commutability of the EQA samples with serum samples of healthy subjects and patients were previously verified [5]. EQA samples were $\log_{10}$-transformed TSH and for each immunoassay method RCALs were applied and the distributions are shown in Supplementary Figure 2. Descriptive statistics were reported in Table 3. Statistics demonstrated that significant differences remain between Access and Architect (Wilcoxon signed rank test, $V = 457$, $p < 0.001$), between Access and Elecsys (Wilcoxon signed rank test, $V = 427$, $p < 0.001$) but not between Architect and Elecsys (Wilcoxon signed rank test, $V = 3869$, p-value = 0.4286). Data were then back-transformed to the original scale (mIU/L). The coefficient of variation (CV) of TSH results of the three immunoassay methods were evaluated both before and after RCAL. The CVs, reported in Figure 1, show that EQA samples were more comparable after applying RCAL than before, the CV being lowered by applying RCAL. The medians (and IQR) of the CVs were 10.72% (6.36%–13.48%) before and 8.16% (6.08% and 11.05%) after recalibration, this difference being highly significant (Wilcoxon signed rant test, $V = 5131$, $p < 0.001$) (in Figure 1 dotted lines represent median CVs).

Recalibration equations are valuable tools for achieving harmonization of TSH results

A second dataset of TSH results, derived from a previous work, was used to test the validity of RCAL regressions [5]. In this dataset, a series of samples were measured

Table 2: Results of recalibration equations (RCALs) for the three TSH immunoassay methods obtained by using bootstrap resampling with M = 1000 iterations and the percentile transformations from the clinical laboratories of the seven Italian city hospitals.

Method	Intercept	95% CI of intercept	Slope	95% CI of slope
Access	−0.00288	−0.01045 to 0.00435	0.98386	0.96961–0.99890
Architect	0.03185	0.02570–0.03789	1.04104	1.02439–1.05714
Elecsys	−0.03121	−0.04024 to −0.02224	1.00304	0.98617–1.01977

simultaneously by all the three immunoassay methods, providing a total of 610, 638 and 639 results from Access, Architect and Elecsys, respectively. After data $\log_{10}$ transformation, RCALs were applied. For each method a further step of inverse recalibration was used to obtain results calibrated against the other two methods. For example, RCALs were used to recalibrate Access data to Architect or Elecsys. The high number of TSH values of this dataset with respect to EQA samples allowed a better comparison of RCAL performances. Moreover, TSH results of this second dataset represent real distributions of TSH values (mIU/L), measured in the Italian population, similar to data used to derive the RCALs.

Comparative values for Access, Architect and Elecsys obtained before (original data) and after RCAL are reported in Table 4. For each method, the recalibration accuracy was evaluated by using mean (±SD) and median and IQR of the differences in the results. Furthermore, p-values of paired comparisons were calculated and reported.

Estimation of TSH reference intervals after RCAL, according to age classes and gender

A published dataset, made of a series of n = 142,821 TSH values stored in the LIS of four Italian city hospitals' clinical laboratories, was used to estimate TSH RI after RCAL. All laboratories used the same analytical method for TSH determination, the Access. Original data (before RCAL) showed a median TSH value equal to 1.75, with 2.5th and 97.5th percentiles being 0.36 and 5.28, respectively [29]. TSH distributions, before and after RCAL, resulted in statistically significant differences (Wilcoxon rank sum test, $W = 1 \times 10\ e^{10}$, $p < 0.0001$), being the 97.5th percentiles 5.28 mIU/L and 5.11 mIU/L, respectively. Considering gender, statistically significant differences were found between the female and male TSH values after RCAL ($X^2 = 465.01$, $p < 0.001$) (Figure 2). Thus, the following age classes were considered: (a) age <35 years, (b) 35 ≤age <50 years, (c) 50 ≤ age < 70 years and (d) age ≥70 years. TSH values after RCAL were statistically significant different among age classes ($X^2 = 1038.58$, Bonferroni's adjusted p-value <0.001 in all comparisons), even after applying a further stratification for both females ($X^2 = 439.34$, Bonferroni's adjusted p-value <0.001 for all comparisons) and males ($X^2 = 722.23$, Bonferroni's adjusted p-value <0.001 for all comparisons). Table 5 reports the median, IQR, RI 2.5th and 97.5th percentiles (and the corresponding 90% CI) of RCAL TSH values, overall, stratified by gender or by gender and age classes. Comparison of percentiles showed that in the female group, 2.5th and the 97.5th

Table 3: Descriptive statistics obtained for EQA $\log_{10}$-transformed TSH data before and after recalibration.

Method	Before applying RCAL			After applying RCAL		
	Median (IQR)	Mean (95% CI)	SD	Median (IQR)	Mean (95% CI)	SD
Access	0.4637 (−0.1385 to 0.9541)	0.3741 (0.2416–0.5068)	0.7299	0.4533 (−0.1392 to 0.9358)	0.3652 (0.2348–0.4956)	0.7182
Architect	0.4316 (−0.1714 to 0.9036)	0.3640 (0.2352–30.4927)	0.7091	0.4812 (−0.1464 to 0.9725)	0.4107 (0.2767–0.5448)	0.7383
Elecsys	0.5335 (−0.0762 to 1.0017)	0.4390 (0.3138–60.5643)	0.6901	0.5039 (−0.1076 to 0.9735)	0.4092 (0.2835–0.5348)	0.6922

Results are expressed as $\log_{10}$ (mIU/L). RCAL, recalibration equation.

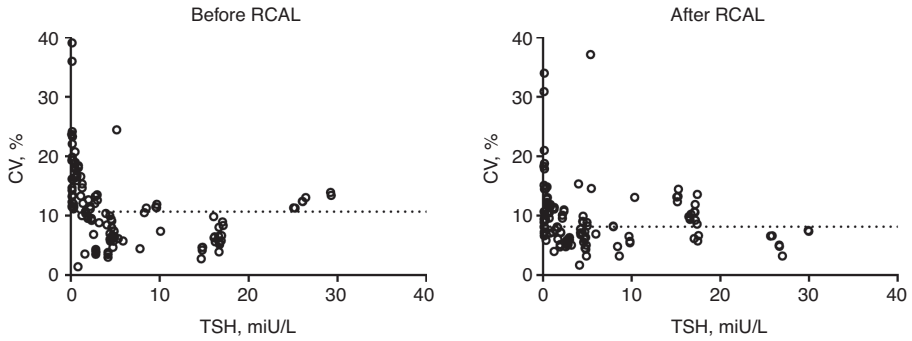

Figure 1: Coefficient of variation in percentage (CV, %) of EQA sample results obtained from three different immunoassays, plotted against the mean TSH value (mIU/L).
Dotted line shows the median CVs % (10.72% before and 8.61% after RCAL).

percentiles differ across age classes, except for the comparison of the 2.5th percentile between groups (a) vs. (c) (pairwise permutation tests across groups for percentiles, p-adjusted = 0.287). In the male groups, significant differences were found across age classes for the 97.5th percentile, while in the case of 2.5th percentiles, statistically significant differences were found for the comparisons between groups (a) vs. (d), (b) vs. (d) and (c) vs. (d).

Discussion

Thyroid dysfunction represents an important endocrine disorder, with a prevalence of 3.82% of cases in Europe [34] and in the clinical setting, TSH laboratory testing is considered key to attaining quality medical care in patients with suspected disease. Indeed, TSH has been recently included in the list of the Top 25 Laboratory Tests by Volume and Revenue [35].

In this study, we investigated and validated a novel statistical approach for harmonizing TSH results, using a workflow based on a two-phase approach. In the first phase, RCALs were estimated using TSH values of a huge explorative dataset obtained from LIS of seven different laboratories of Italy, and a validation phase, performed using two different testing datasets.

The basic assumption of the approach proposed in this study is that TSH values obtained from different immunoassays represent subsamples of the whole distribution of TSH, the population data. Considering that each method measures subsamples originating from the same distribution, harmonization can be obtained by recalibrating through merged data (randomly resampled) from different immunoassays. The percentile transformation implemented in this study is similar to the "percentile normalization" approach, used for normalizing other types of data (e.g. microarray data), a technique for making two distributions similar for statistical properties. Interestingly, this is a distributional "constrains-free" approach and does not suffer from the limitations that are typical of parametric methods, such as deviation from normality and outliers. To our knowledge, the usage of bootstrapping resampling and of percentile transformation represents a novel approach for harmonizing immunoassay distributions of TSH values.

In the first phase, the dataset used included a series of 125,419 TSH values and after outlier removal, average and median values of the three distributions were different. For Access and Architect, the estimated RIs were wider than the manufacturers' declared values, while for Elecsys they were closer than RIs declared into the insert. However, TSH results being highly skewed, data were $\log_{10}$ transformed before further analyses. Shapes of method distributions and of the overall distribution (made from all method results) deviated from the normal, especially for the large left and thick right tails, which might underline that subjects with normal TSH, but with values at the boundary or exceeding the range 2.5th and 95th percentiles, are relevantly under- or overestimated by the different methods, especially for the Architect results (Supplementary Figure 1). Percentile transformations were applied to transform data. As stated above, the applicability of this method is not limited by deviation from the log-normal TSH distributions.

RCAL results are obtained by implementing the robust regression (Huber method) with bootstrap resampling. Regression intercepts and slopes, calculated by averaging the bootstrapped results of several iterations, presented different meanings: the first parameter is used to correct for a systematic component across methods; in contrast, slopes account for random, method-dependent, variations.

Table 4: Harmonization results obtained by applying RCAL.

Immunoassay	Differences in mIU/L with respect to:					
	Architect recalibrated to Access			Elecsys recalibrated to Access		
Access	**Mean of differences (±SD)**	**Median of differences (IQR)**	**p-Value**	**Mean of differences**	**Median (IQR)**	**p-Value**
	0.1955 (±6.1847)	−0.1587 (−0.5975 to −0.0040)	<0.001	0.1954 (±3.3324)	−0.0053 (−0.2859 to 0.3082)	0.955
	Access recalibrated to Architect			Elecsys recalibrated to Architect		
Architect	**Mean of differences (SD)**	**Median (IQR)**	**p-Value**	**Mean of differences**	**Median (IQR)**	**p-Value**
	−0.08697 (±4.1676)	0.1388 (0.0041–0.4892)	<0.001	0.0393 (±2.9450)	0.1380 (0.0129–0.3678)	<0.001
	Architect recalibrated to Elecsys			Access recalibrated to Elecsys		
Elecsys	**Mean of differences (SD)**	**Median (IQR)**	**p-Value**	**Mean of differences**	**Median (IQR)**	**p-Value**
	−0.1863 (±3.2420)	0.0060 (−0.3106 to 0.2987)	0.9724	−0.0154 (±4.1790)	−0.1643 (−0.4625 to −0.0149)	<0.001

The comparison between original TSH results (mIU/L) obtained by immunoassays before and after applying recalibration equations (RCALs) were reported.

Figure 2: Scatterplot of $\log_{10}$ TSH values (obtained after RCAL) and age.
Cubic spline was used to obtain the TSH trend for female (red) and male (blue) subjects.

Confidence intervals obtained by bootstrap statistics showed that for Access, intercept was not statistically significant, while for Elecsys, slope was not significant.

RCAL equations were validated by using two other datasets. Firstly, the EQA results were used to assess the validity of RCAL equations. The comparison of EQA distributions of TSH values before and after applying RCALs on $\log_{10}$-transformed values showed a good agreement among results, the median, 25th and 75th percentiles values being more similar (Table 3 and Supplementary Figure 2). Also, the variability of the results obtainable by measuring the same EQA samples with the three methods are more comparable, being the CV values highly significantly reduced after RCAL (Figure 1).

A second dataset, containing numerous TSH results obtained by simultaneous measurements using the Access, Architect and Elecsys methods, was further used to verify the harmonization properties of RCAL equations. Each method was recalibrated against the other two, in order to achieve a direct method-to-method comparison. The results reported in Table 4 showed that, after recalibration, agreements were very good across all methods, the median of differences ranging from −0.0053 to 0.1955 mIU/L of TSH. Statistical testing of paired data showed that results from Elecsys recalibrated to Access and vice-versa belong to the same data distribution. These results demonstrated that distributions of TSH results for different immunoassay methods can be harmonized to originating comparable results.

A further huge dataset, derived from TSH stored in the LIS of four Italian city hospitals' clinical laboratories, was used to estimate RIs of TSH after RCAL. Results

Table 5: Descriptive statistics, 2.5th and 97.5th percentiles (and the corresponding 90% CI) of data distribution of TSH values (mIU/L) measured by different laboratories with the method Access after recalibration equations (RCALs).

Access	Median (IRQ)	2.5th (90% CI)	97.5th (90% CI)
After RCAL			
Overall data	1.72 (1.14–2.54)	0.37 (0.36–0.37)	5.11 (5.06–5.13)
Female	1.77 (1.16–2.60)	0.35 (0.34–0.36)	5.16 (5.29–5.39)
Male	1.62 (1.09–2.37)	0.40 (0.39–0.41)	4.92 (4.83–4.98)
Females			
Age <35	1.87 (1.31–2.64)	0.49 (0.46–0.51)	4.96 (4.87–5.05)
Age ≥35–50	1.77 (1.21–2.53)	0.38 (0.37–0.40)	4.81 (4.73–4.92)
Age ≥50–70	1.76 (1.14–2.61)	0.33 (0.32–0.34)	5.13 (5.07–5.21)
Age ≥70	1.70 (1.05–2.63)	0.31 (0.29–0.32)	5.48 (5.41–5.54)
Males			
Age <35	1.96 (1.41–4.88)	0.66 (0.64–0.70)	4.88 (4.74–5.05)
Age ≥35–50	1.62 (1.14–2.30)	0.49 (0.47–0.51)	4.65 (4.52–4.73)
Age ≥50–70	1.55 (0.39–2.27)	0.39 (0.37–0.41)	4.73 (4.63–4.83)
Age ≥70	1.54 (0.99–5.27)	0.35 (0.33–0.36)	5.27 (5.18–5.36)

Data considered overall or subdivided by females and males, with or without stratification for age-classes, are reported. Bootstrap statistics (with R = 1000 iterations) was used to derive 90%CI of the 2.5th and 97.5th percentiles.

showed that before and after RCAL results lead to different RIs, especially considering the upper limit and that TSH values are age and gender correlated (Figure 2). Therefore, age- and gender-specific RIs were calculated (Table 5). Especially for the lower RI limit, statistically significant differences were observed stratifying the TSH values for age and gender. Interestingly, while the upper bound of age-specific RIs showed a "U-shape" trend, the lower bound decreased with increasing age. These results shows a decreasing trend of TSH with age, and are in accordance with previously published results of the same group [29], while in the study by Lo Sasso et al., made in the same geographical area and with a different analytical method, the "U-shape" was less remarked in the upper RI limit. In the latter study, RIs were 0.18 mIU/L (90% CI 0.14–0.21) and 3.54 mIU/L (90% CI 3.18–3.90), underlining a marked difference, especially in the upper RI limit [36].

Previous studies from Clerico et al. [4] and Stöckl et al. [14] used other statistical approaches for harmonizing TSH results, reporting successful results. Both studies used robust factor analyses and are based on TSH results obtained from the measurement of a relatively low number of clinical and/or quality-control samples with some TSH methods. In contrast, in the present study, huge datasets were used both to derive RCALs and to validate their efficacy in harmonizing TSH distributions. Two-phase approaches allow reducing "overoptimistic" results that can be obtained sometimes when single-step methods are used. Further, the usage of the large dataset increases the generalizability of RCAL estimations, increasing the applicability of the obtained results [37]. Other advantages with respect to previous TSH harmonization attempts [4, 14] are represented by the combination of percentile transformation with robust regression for estimating the RCAL, and the usage of bootstrapping for iterating the whole procedure. On the one hand this allow to reduce the influence of large left and thick right tails on RCAL estimations. On the other hand, during bootstrapping, subsamples of the same size are considered and merged several times, further limiting the effects of outlier TSH results [30].

The aspect of accounting interferences is of utmost importance for the harmonization of TSH results [8, 9]. Potential interferences in immunometric methods with rheumatoid factors or heterophilic antibodies can occasionally cause abnormal TSH concentrations [38, 39]. Further, samples with elevated levels of biotin or anti-streptavidin antibodies might produce variability between analyzers [38, 39]. Interferences in immunometric methods could be considered as outliers [8, 9], and their frequency for TSH testing is estimated to be up to 1% [38]. Accordingly, the robust statistical approach used in this study should also be able to eliminate the interferences, which produce TSH values above or below the reference range (i.e. abnormal values).

The major disadvantage of the approach proposed in this study with respect to that proposed by Clerico et al. [4] and Stöckl et al. [14] is the high complexity of statistics due to the bootstrapping application, even if the usage of these resampling methods is limited to the initial setup of RCAL.

In conclusion, the present study represents insights on the utilization of the percentile transformation approach and of bootstrapping statistics to derive RCALs, which are useful to harmonize TSH values. Further, this approach can lead to the estimation of method-independent RIs for TSH. On the other hand, this approach also presents some important limitations. For example, robust and valuable RCAL estimations require the availability of big data sources, collected possibly from different laboratories and in different time periods, which complicated the applicability of bootstrapping methods that are computationally intensive.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.
Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Spencer CA. Assay of thyroid hormones and related substances. South Dartmouth, MA, USA: Endotext Inc., 2000 MDText.com. https://www.ncbi.nlm.nih.gov/books/NBK279113/. Accessed: Dec 2019.
2. Garber J, Cobin R, Gharib H, Hennessey J, Klein I, Mechanick J, et al. Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association. Endocr Pract 2012;18:988–1028.
3. Alexander EK, Pearce EN, Brent GA, Brown RS, Chen H, Dosiou C, et al. Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and the postpartum. Thyroid 2017;27:315–89.
4. Clerico A, Ripoli A, Zucchelli GC, Plebani M. Harmonization protocols for thyroid stimulating hormone (TSH) immunoassays: different approaches based on the consensus mean value. Clin Chem Lab Med 2015;53:377–82.
5. Clerico A, Ripoli A, Fortunato A, Alfano A, Carrozza C, Correale M, et al. Harmonization protocols for TSH immunoassays: a multicenter study in Italy. Clin Chem Lab Med 2017;55:1722–33.
6. Iervasi G, Clerico A, Bonini R, Manfredi C, Berti S, Ravani M, et al. Acute effects of amiodarone administration on thyroid function in patients with cardiac arrhythmia. J Clin Endocrinol Metab 1997;82:275–80.
7. Verdickt L, Maiter D, Depraetere L, Gruson D. TSH-assay interference: still with us. Clin Lab 2012;58:1305–7.
8. Clerico A, Plebani M. Biotin interference on immunoassay methods: sporadic cases or hidden epidemic? Clin Chem Lab Med 2017;55:777–9.
9. Clerico A, Belloni L, Carrozza C, Correale M, Dittadi R, Dotti C, et al. A Black Swan in clinical laboratory practice: the analytical error due to interferences in immunoassay methods. Clin Chem Lab Med 2018;56:397–402.
10. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 1: thyroid-stimulating hormone. Clin Chem 2010;56:902–11.
11. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 2: free thyroxine and free triiodothyronine. Clin Chem 2010;56:912–20.
12. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 3: total thyroxine and total triiodothyronine. Clin Chem 2010;56:921–9.
13. Thienpont LM, Van Uytfanghe K, Van Houcke S, Das B, Faix JD, MacKenzie F, et al. A progress report of the IFCC Committee for Standardization of Thyroid Function Tests. Eur Thyroid J 2014;3:109–16.
14. Stöckl D, Van Uytfanghe K, Van Aelst S, Thienpont LM. A statistical basis for harmonization of thyroid stimulating hormone immunoassays using a robust factor analysis model. Clin Chem Lab Med 2014;52:965–72.
15. Gantzer ML, Miller WG. Harmonisation of measurement procedures: how do we get it done? Clin Biochem Rev 2012;33:95–100.
16. Van Houcke SK, Thienpont LM. "Good samples make good assays" – the problem of sourcing clinical samples for a standardization project. Clin Chem Lab Med 2013;51:967–72.
17. Van Houcke SK, Van Aelst S, Van Uytfanghe K, Thienpont LM. Harmonization of immunoassays to the all-procedure trimmed mean – proof of concept by use of data from the insulin standardization project. Clin Chem Lab Med 2013;51:e103–5.
18. Van Uytfanghe K, De Grande LA, Thienpont LM. A "Step-Up" approach for harmonization. Clin Chim Acta 2014;432:62–7.
19. Plebani M. Harmonization in laboratory medicine: the complete picture. Clin Chem Lab Med 2013;51:741–51.
20. Plebani M, Panteghini M. Promoting clinical and laboratory interaction by harmonization. Clin Chim Acta 2014;432:15–21.
21. Plebani M, Astion ML, Barth JH, Chen W, de Oliveira Galoro CA, Escuer MI, et al. Harmonization of quality indicators in laboratory medicine. A preliminary consensus. Clin Chem Lab Med 2014;52:951–8.
22. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. Clin Chim Acta 2003;327:25–37.
23. Greene DN, Leong TK, Collinson PO, Kamer SM, Huang K, Lorey TS, et al. Age, sex, and racial influences on the Beckman Coulter AccuTnI+3 99th percentile. Clin Chim Acta 2015;444:149–53.
24. Baloch Z, Carayon P, Conte-Devolx B, Demers LM, Feldt-Rasmussen U, Henry JF. Laboratory support for the diagnosis and monitoring of thyroid disease. Thyroid 2003;13:3.
25. Arzideh F, Wosniok W, Haeckel R. Indirect reference intervals of plasma and serum thyrotropin (TSH) concentrations from intralaboratory data bases from several German and Italian medical centres. Clin Chem Lab Med 2011;49:659–64.

26. Vadiveloo T, Donnan PT, Murphy MJ, Leese GP. Age- and gender-specific TSH reference intervals in people with no obvious thyroid disease in Tayside, Scotland: the Thyroid Epidemiology, Audit, and Research Study (TEARS). J Clin Endocrinol Metab 2013;98:1147–53.
27. Farrell C-J, Nguyen L, Carter AC. Data mining for age-related TSH reference intervals in adulthood. Clin Chem Lab Med 2017;55:e213–5.
28. Tozzoli R, D'Aurizio F, Metus P, Steffan A, Mazzon C, Bagnasco M. Reference intervals for thyrotropin in an area of Northern Italy: the Pordenone thyroid study (TRIPP). J Endocrinol Invest 2018;41:985–94.
29. 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 2018;57:259–67.
30. Baadenhuijsen H, Smit JC. Indirect estimation of clinical chemical reference intervals from total hospital patient data: application of a modified Bhattacharya procedure. J Clin Chem Clin Biochem 1985;23:829–39.
31. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141–64.
32. Huber PJ, Ronchetti E. Robust statistics, 2nd ed. Hoboken, NJ, USA: Wiley, 2009.
33. Simundic A-M. Practical recommendations for statistical analysis and data presentation in Biochemia Medica journal. Biochem Medica 2012;22:15–23.
34. Garmendia Madariaga A, Santos Palacios S, Guillén-Grima F, Galofré JC. The incidence and prevalence of thyroid dysfunction in Europe: a meta-analysis. J Clin Endocrinol Metab 2014;99:923–31.
35. Horton S, Fleming KA, Kuti M, Looi L-M, Pai SA, Sayed S, et al. The top 25 laboratory tests by volume and revenue in five different countries. Am J Clin Pathol 2019;151:446–51.
36. Lo Sasso B, Vidali M, Scazzone C, Agnello L, Ciaccio M. Reference interval by the indirect approach of serum thyrotropin (TSH) in a Mediterranean adult population and the association with age and gender. Clin Chem Lab Med 2019;57:1587–94.
37. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309–14.
38. Favresse J, Burlacu M-C, Maiter D, Gruson D. Interferences with thyroid function immunoassays: clinical implications and detection algorithm. Endocr Rev 2018;39:830–50.
39. Thayakaran R, Adderley NJ, Sainsbury C, Torlinska B, Boelaert K, Šumilo D, et al. Thyroid replacement therapy, thyroid stimulating hormone concentrations, and long term health outcomes in patients with hypothyroidism: longitudinal study. Br Med J 2019;366.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2019-1167).