Generating method-specific Reference Ranges – A harmonious outcome?

CrossMark

Graham R. Lee*, Alison Griffin, Kieran Halton, Maria C. Fitzgibbon

Department of Clinical Biochemistry and Diagnostic Endocrinology, Mater Misericordiae University Hospital, Eccles Street, Dublin, D7, Ireland

ARTICLE INFO

ABSTRACT

Objectives: When laboratory Reference Ranges (RR) do not reflect analytical methodology, result interpretation can cause misclassification of patients and inappropriate management. This can be mitigated by determining and implementing method-specific RRs, which was the main objective of this study.

Design and methods: Serum was obtained from healthy volunteers (Male + Female, n > 120) attending hospital health-check sessions during June and July 2011. Pseudo-anonymised aliquots were stored (at − 70 °C) prior t° analysis on Abbott ARCHITECT c16000 chemistry and *i*2000SR immunoassay analysers. Data were stratified by gender where appropriate. Outliers were excluded statistically (Tukey method) to generate non-parametric RRs (2.5th + 97.5th percentiles). RRs were compared to those quoted by Abbott and UK Pathology Harmony (PH) where possible. For 7 selected tests, RRs were verified using a data mining approach.

Results: For chemistry tests (n = 23), Upper or Lower Reference Limits (LRL or URL) were > 20% different from Abbott ranges in 25% of tests (11% from PH ranges) but in 38% for immunoassay tests (n = 13). RRs (mmol/L) for sodium (138−144), potassium (3.8–4.9) and chloride (102−110) were considerably narrower than PH ranges (133–146, 3.5–5.0 and 95–108, respectively). The gender difference for ferritin (M: 29–441, F: 8–193 ng/mL) was more pronounced than reported by Abbott (M: 22–275, F: 5–204 ng/mL). Verification studies showed good agreement for chemistry tests (mean [SD] difference = 0.4% [1.2%]) but less so for immunoassay tests (27% [29%]), particularly for TSH (LRL).

Conclusion: Where resource permits, we advocate using method-specific RRs in preference to other sources, particularly where method bias and lack of standardisation limits RR transferability and harmonisation.

1. Introduction

Reference Ranges (RRs) are intended to guide clinicians on the interpretation of a patient's test results, in the context of the patient's overall clinical assessment [1]. RRs are a familiar component of the result report and may have particular value when there is no previous patient data for comparison. However, their statistical derivation, inherent uncertainty and limitations are unlikely to be appreciated for routine test result evaluations, particularly by less experienced clinicians. For some analytes, e.g. Cholesterol and

* Corresponding author.
E-mail address: glee@mater.ie (G.R. Lee).

HbA1C, RRs have been replaced by decision limits and therapeutic targets and so some of the laboratory's focus on reducing error is diverted to other extra-analytical and analytical stages of the Total Testing Pathway [2]. Where a patient has serial results available, RRs are believed to have less value in monitoring patient status [3]. Instead, an evaluation of the magnitude of any result change, as compared to a test's Reference Change Value (RCV), may provide a better means for monitoring patients and guiding management decisions [3]. RCVs are particularly useful for highlighting significant changes even when patient results remain within the RR and thereby may provide an earlier indication of changing patient status. Furthermore, use of RCVs may provide appropriate context particularly when results change from inside to outside the RR (and *vice versa*) but without significance (change < RCV).

Previous work by Harris [4] showed that the usefulness of RRs in evaluating a patient's results is limited to a minority of analytes showing a large Intra- to Inter-individual Coefficient of Variation ratio ($CV_I:CV_G > 1.4$). For such analytes, serial results for a (stable) patient may potentially span the entire RR rather than only a limited part of it, and therefore the RR can be a useful and sensitive tool for assessing serial changes [4,5]. RRs also permit more meaningful result interpretation when they are relevant not only to the patient's demographics (e.g. race, gender) but also to the analytical methodology generating the test result [6]. The latter factor is particularly important when large inter-methodological bias exists. Careful consideration must therefore be given to defining, establishing and verifying RRs. For such purposes, the International Federation for Clinical Chemistry and Laboratory Medicine (IFCC) and the Clinical Laboratory Standards Institute (CLSI) offer guidance and specific recommendation for non-parametric RRs, derived from a minimum of 120 reference individuals [7]. Alternative approaches such as data mining of patient data are practically advantageous, less resource intensive and can provide large (and representative) data for determining RRs with a high level of confidence [1]. Where assay standardisation has assisted comparability of results between different analytical methods, a universal or harmonised RR may also be used preferentially [6]. The transferability of such RRs across healthcare systems is welcomed by clinicians, especially when movement of patients and clinicians within or between networked hospitals results in the same test being performed with different laboratory methods. This is the "pragmatic science" approach adopted by the United Kingdom (UK) Pathology Harmony (PH) group [8], whose aim is to develop a common set of RRs on the premise that minor methodological differences between laboratories are of little clinical significance. However, this approach is not currently viable for many immunoassay based tests e.g. troponins and tumour markers, whose assays are not standardized, with concomitant analytical bias between method manufacturers [9,10].

When appropriate RRs are however established, regardless of the approach, it is also important to periodically verify their use and applicability to the current measurement system [8,11]. This is in accord with the International Organisation for Standardisation's Requirements for Quality and Competence in Medical Laboratories (ISO15189:2012), which recommends verification following a change to the examination or pre-examination procedure [12].

In this study, we report on our own experiences and challenges in establishing method-specific RRs across our clinical biochemistry and diagnostic endocrinology test repertoire during a period of laboratory consolidation and verification of new analytical platforms.

2. Materials and methods

2.1. Reference Range subjects

This RR study was part of service evaluation, involving surplus sera from adult volunteers (non-fasted) who had previously attended hospital health check sessions in June and July 2011. Only sera from volunteers less than 18 years of age were excluded from analysis. There were no other exclusion criteria.

2.2. Pre-analytical: collection, archiving and preparation of serum samples

Blood from volunteers (10 mL) was initially collected into serum separator tubes (Becton Dickinson, Franklin Lakes, NJ, USA) and centrifuged at 10,000*g*, for 10 min at room temperature prior to analysis for "health check" purposes. Post-analysis, aliquots of residual serum (1.5 mL) were pseudo-anonymised (coded) and stored (at − 70 °C) until analysis for RR purposes. On the day of analysis, stored serum samples were thawed slowly while rotating at r.t.p. before transfer to 1.5 mL plastic tubes (Sarstedt, Numbrecht, Germany).

2.3. Analytical

All reagents and calibrators were obtained from Abbott Diagnostics (Abbott Laboratories, Lake Forest, IL, USA). Internal Quality Control (IQC) material was obtained from Technopath (Ballina, Ireland) and was run at 2 or 3 levels for each analyte. Running imprecision data (CVs) are shown in Table 1. Each test was also assessed through participation in relevant schemes run by the United Kingdom National External Quality Assurance Service (UKNEQAS, Birmingham, UK).

Serum samples (n > 120) were analysed over 5 days on Abbott ARCHITECT *c*16000 and *i*2000SR analysers for a comprehensive list of chemistry (n = 23) and immunoassay (n = 13) tests. All maintenance, calibration and IQC procedures were followed in accordance with the manufacturer's instructions.

Table 1

Internal Quality Control (IQC). IQC running data (5 days, n = 25) are shown for all chemistry (top) and immunoassay (bottom) based tests. L = Level, SD = Standard Deviation, CV = Coefficient of Variation.

Test	ALT (U/L)			AST (U/L)			Albumin (BCG, g/L)			Albumin (BCG, g/L)			ALP (U/L)		
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Mean	24.4	93.0	204	33.5	110	238	24.2	39.8	50.4	24.2	38.4	48.2	53.6	142	293
SD	0.82	7.39	5.39	0.51	1.08	1.26	0.21	0.27	0.29	0.15	0.17	0.29	0.87	1.47	2.24
CV	3.4	8.0	2.6	1.5	1.0	0.5	0.9	0.7	0.6	0.6	0.5	0.6	1.6	1.0	0.8

Test	Amylase (U/L)			Bilirubin (Direct, μmol/L)			Bilirubin (Total, μmol/L)			Calcium (mmol/L)			CK (U/L)		
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Mean	45.1	145	413	8.0	26.0	49.3	16.5	56.6	105	1.60	2.55	3.26	75.0	212	430
SD	0.33	0.96	3.22	0.09	0.60	0.40	0.25	1.12	1.31	0.01	0.02	0.02	0.84	2.37	6.14
CV	0.7	0.7	0.8	1.1	2.3	0.8	1.5	2.0	1.2	0.8	0.9	0.5	1.1	1.1	1.4

Test	Chloride (mmol/L)			Creatinine (Enz., μmol/L)			CRP (mg/L)			GGT (U/L)			Iron (μmol/L)		
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Mean	83.0	95.8	113	50.2	191	557	1.0	2.3	46.0	20.6	73.6	152	13.2	23.2	31.7
SD	0.00	0.37	0.46	0.75	1.70	3.31	0.03	0.04	0.41	0.57	0.71	0.86	0.07	0.13	0.15
CV	0.0	0.4	0.4	1.5	0.9	0.6	3.0	1.9	0.9	2.8	1.0	0.6	0.5	0.5	0.5

Test	LDH (U/L)			Magnesium (mmol/L)			Phosphate (mmol/L)			Potassium (mmol/L)			Protein (Total, g/L)		
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Mean	165	297	580	0.41	0.98	1.61	0.68	1.20	2.47	2.50	4.08	7.18	37.6	62.4	81.9
SD	2.79	3.27	5.80	0.02	0.02	0.02	0.01	0.03	0.02	0.00	0.04	0.05	0.16	0.27	0.30
CV	1.7	1.1	1.0	4.0	1.7	1.0	1.2	2.5	0.9	0.0	1.1	0.7	0.4	0.4	0.4

Test	Sodium (mmol/L)			Transferrin (g/L)			Urea (mmol/L)			Uric Acid (μmol/L					
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3			
Mean	115	138	161	1.5	2.5	3.2	4.8	13.7	22.8	191	381	505			
SD	0.51	0.54	0.48	0.01	0.01	0.01	0.08	0.10	0.17	1.99	3.66	5.99			
CV	0.4	0.4	0.3	0.8	0.5	0.5	1.6	0.7	0.7	1.0	1.0	1.2			

Test	Anti-Tg (IU/mLl)			Anti-TPO (IU/mL)			B12 (ng/L)			Cortisol (nmol/L)			Ferritin (μg/L)		
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Mean	25.2	21.0	25.1	22.6	32.2	39.8	310	733	1187	59.7	370	680	20.7	301	709
SD	1.1	0.8	2.4	0.8	0.8	0.9	14.1	18.9	55.0	2.7	9.7	17.7	0.6	9.4	29.3
CV	4.3	3.6	9.6	3.6	2.4	2.2	4.6	2.6	4.6	4.5	2.6	2.6	2.9	3.1	4.1

Test	Folate (μg/L)			FSH (mIU/mL)			free T4 (pmol/L)			free T3 (pmol/L)			LH (mIU/mL)		
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Mean	2.9	5.7	NA	3.8	19.9	42.8	6.5	28.9	60.3	2.0	6.4	11.6	3.8	17.3	45.2
SD	0.2	0.2	NA	0.1	0.6	1.2	0.4	0.5	3.6	0.2	0.2	0.3	0.1	0.7	1.6
CV	7.7	3.9	NA	2.4	2.9	2.9	5.6	1.6	6.0	11.6	3.0	2.7	2.4	4.1	3.4

Test	Prolactin (mIU/L)			PTH (pmol/L)			TSH (mIU/mL)								
IQC Level	L1	L2	L3	L1	L2	L3	L1	L2	L3						
Mean	125	308	2373	23.3	43.1	172	0.1	5.1	30.5						
SD	2.0	5.5	57.0	0.8	2.3	7.1	0.0	0.1	0.9						
CV	1.6	1.8	2.4	3.3	5.3	4.2	2.7	1.9	2.9						

2.4. Post-analytical

2.4.1. Retrospective RR verification

Method-specific RRs were verified subsequently for a selection of analytes [Albumin, alkaline phosphates (ALP), Calcium (+adjusted), Magnesium, Phosphate, Ferritin and thyroid stimulating hormone (TSH)]. These were selected after discussion of our RRs (as determined above) with clinical users and/or where marked differences (> 20%) to Abbott's quoted RRs were observed. For these 7 analytes, our Laboratory Information System (LIS) was interrogated for results from GP patients (September to October 2011). Only patients 18 years and older were included. Patients were excluded if their creatinine levels were outside our method- and gender-specific RRs.

2.4.2. Data analysis and statistics

Reference Ranges were determined for all analytes, including those selected for RR verification studies, using Medcalc statistical software, version 14 (Medcalc, Ostend, Belgium). Using this software data outliers were initially detected statistically by the Tukey method [7]. whereby 'Far Out' [FO] (and 'Outside' [OS]) outliers were defined as values which were either smaller than the lower quartile minus 3 (1.5 for OS) times the inter-quartile range (IQR) or larger than the upper quartile plus 3 (1.5 for OS) times the IQR. After excluding FO outliers, Medcalc was then used to estimate non-parametric 95% RRs and 90% Confidence Intervals (CI) for each analyte. Although the age and gender of each volunteer was known, gender-specific RRs were only determined for Creatine Kinase (CK), Creatinine, Gamma Glutamyl Transferase (GGT), Iron, Transferrin, Ferritin and Prolactin. For women, surplus sera were available from only 118 of the volunteers during the period of heath check testing, therefore female RRs for the above seven tests (n

Table 2

Reference Ranges (RR) for Chemistry tests. Upper and Lower Reference Limits (URLs and LRLs) quoted by Abbott and Pathology Harmony (PH) and those determined in the current study (MMUH – Mater Misericordiae University Hospital) are shown. Female-specific RRs (+ CIs, n = 100–118), as determined using the robust method [7] (where n < 120) are shown in italics. Differences (%) between Abbott and PH URLs/LRLs versus respective MMUH limits are also shown. *Tests where Abbott-based RRs are quoted – see Discussion. PP = Pyridoxal Phosphate, BCG/P = Bromocresol Green/Purple, Enz. = Enzymatic, NA = Not Available, CI = Confidence Interval.

Test	Abbott		PH		MMUH				MMUH vs Abbott		MMUH vs PH	
	LRL	URL	LRL	URL	LRL	90% CI	URL	90% CI	LRL	URL	LRL	URL
ALT (+ PP), (U/L)	0	55	NA	NA	10.0	7-12	59.0	46-76	NA	6.8	NA	NA
AST, (U/L)	5	34	NA	NA	18.0	14-20	44.0	39-64	72.2	22.7	NA	NA
Albumin (BCG, g/L)	35	50	35	50	38.0	37.4-38.5	45.0	44.3-46.7	7.9	-11.1	7.9	-11.1
Albumin (BCP, g/L)	35	50	35	50	36.0	34.3-36.4	44.0	43.4-44.9	2.8	-13.6	2.8	-13.6
ALP (U/L)	40	150	30	130	38.0	32-46	108.0	95-148	-5.3	-38.9	21.1	-20.4
*Amylase (U/L)	23	96	NA	NA	26.0	18-29	98.0	90-131	11.5	2.0	NA	NA
*Bilirubin (Total, μmol/L)	3.4	20.5	NA	21	5.0	4.5-5.9	25.5	19.3-26.6	32.5	19.6	NA	17.6
*Bilirubin (Direct, μmol/L)	0	8.6	NA	NA	2.3	1.9-2.5	8.5	6.2-9.3	NA	-1.2	NA	NA
*Calcium (mmol/L)	2.23	2.5	NA	NA	2.12	2.1-2.2	2.53	2.45-2.58	-5.2	1.2	NA	NA
CK (U/L)												
Female	29	168	25	200	32.0	NA	166.0	148-181	9.4	-1.2	21.9	-20.5
Male	30	200	40	320	44.0	36-53	279.0	230-374	31.8	28.3	9.1	-14.7
*Chloride (mmol/L)	101	110	95	108	102.0	101-102.6	110.0	109.3-111.2	1.0	0.0	6.9	1.8
Creatinine (Enz., μmol/L)												
Female	49	90	NA	NA	46.0	43-48	85.0	82-87	-6.5	-5.9	NA	NA
Male	64	104	NA	NA	66.0	64-69	107.0	106-116	3.0	2.8	NA	NA
CRP (mg/L)	0	5	NA	NA	0.2	0.1-0.2	6.9	4.7-7.8	NA	27.5	NA	NA
GGT (U/L)												
Female	9	36	NA	NA	8.0	NA	38.0	33-42	-12.5	5.3	NA	NA
Male	12	64	NA	NA	11.0	9-14	68.0	54-92	-9.1	5.9	NA	NA
Iron (μmol/L)												
Female	4.5	28	NA	NA	7.0	5.4-8.8	32.0	30.1-33.5	35.7	12.5	NA	NA
Male	5.5	26	NA	NA	12.0	9.7-13	31.8	31.2-33.5	54.2	18.2	NA	NA
*LDH (U/L)	125	243	NA	NA	119.0	101-129	220.0	210-235	-5.0	-10.5	NA	NA
Magnesium (mmol/L)	0.66	1.07	0.7	1	0.79	0.76-0.80	1.01	0.99-1.06	16.5	-5.9	11.4	1.0
Phosphate (mmol/L)	0.74	1.52	0.8	1.5	0.82	0.68-0.87	1.39	1.35-1.44	9.8	-9.4	2.4	-7.9
*Potassium (mmol/L)	3.5	5	3.5	5.3	3.8	3.8-3.9	4.9	4.8-5.4	7.9	-2.0	7.9	-8.2
Protein (Total, g/L)	64	83	60	80	63.0	59-64.1	76.0	75.6-77.2	-1.6	-9.2	4.8	-5.3
*Sodium (mmol/L)	137	144	133	146	138.0	137.3-138.2	144.0	143.3-145.2	0.7	0.0	3.6	-1.4
Transferrin (g/L)												

Difference >20%

Difference 10-20%

Difference <10%

= 101–118 *per* test) were determined using Medcalc by the Robust method [7]. The requirement to partition the data by gender was confirmed using the Harris and Boyd standard deviate test [13].

2.4.3. Biochemical, endocrinological and haematological correlation

For data obtained in the initial RR study, FO and OS outliers were further correlated against available biochemical, endocrinological and haematological (health check) results. In this *a posteriori* approach, OS outliers were also removed if supported by the results of other relevant tests e.g. correlation of transferrin to ferritin and haemoglobin results and correlation of anti-thyroid peroxidase antibodies to fT4 and TSH results, and following independent adjudication by an experienced clinical biochemist.

3. Results

3.1. Subject and patient demographics

3.1.1. Reference Range (RR) study

Serum samples were obtained from 263 volunteers (median [range] age 51 [19–81] years) from which RRs were determined (n = 101–139 per analyte). 55% of subjects were male (median age 51 [19–80] years) and 45% were female (median age 50 [26–81] years).

3.1.2. RR verification study

This study included a pool of 4579 patients from which a number of patients (n = 86–4228) were used to verify the RRs of the seven selected analytes. This pool comprised 38% men (median [range] age 55 [15–96] years) and 62% women (median [range] 53 [14–97] years).

3.2. RR comparisons and evaluations

3.2.1. Chemistry RRs

RRs were determined for 23 chemistry tests (Table 2) and 3 derived calculations (Table 3) and compared to Abbott and PH quoted

Table 3

Reference Ranges (RR) for Calculated tests. Upper and Lower Reference Limits (URLs and LRLs) quoted by Abbott and Pathology Harmony (PH) and those determined in the current study (MMUH – Mater Misericordiae University Hospital) are shown. Differences (%) between Abbott and PH URLs/LRLs versus respective MMUH limits are also shown. Adj. = Adjusted, Sat. = Saturation, CI = Confidence Interval, NA = Not Available. Female-specific RRs (+ CIs, n = 100–118), as determined using the robust method [7] (where n < 120) are shown in italics.

Test	Abbott		PH		MMUH				MMUH vs Abbott		MMUH vs PH	
	LRL	URL	LRL	URL	LRL	90% CI	URL	90% CI	LRL	URL	LRL	URL
Adj. Calcium (mmol/L)	NA	NA	2.2	2.6	2.18	2.16-2.20	2.44	2.42-2.46	NA	NA	-0.9	-6.6
Globulin (g/L)	NA	NA	NA	NA	23	22-25	36	34-38	NA	NA	NA	NA
%Transferrin Sat.												
Female	NA	NA	NA	NA	*10*		*51*		NA	NA	NA	NA
Male	NA	NA	NA	NA	18	16-20	56	53-58	NA	NA	NA	NA

Difference >20%
Difference 10-20%
Difference <10%

RRs, where possible. Across all chemistry tests and for both genders, URLs and LRLs were > 20% different from Abbott in 18% of comparisons (10/56) whereas for the majority (33/56), any discrepancies were < 10%. Differences to PH quoted RRs were also < 10% for the majority of tests (66%, 18/27) and differed by > 20% in only 11% (3/27). RRs for adjusted calcium (AdjCa) differed from PH by < 10% (Table 3). The LRLs and URLs quoted by Abbott and/or PH were outside our RR's 90% CIs for the majority (24/29) of chemistry tests (including AdjCa URL) (Fig. 1).

3.2.2. Immunoassay RRs

Calculated RRs were determined for 13 immunoassay tests and were compared to Abbott's quoted RRs as shown in Table 4. Across all tests and for both genders, calculated limits were > 20% different from Abbott URL and LRLs for 38% of tests (10/26), and were < 10% different in 35% (9/26). The LRLs and URLs quoted by Abbott were outside our RR's 90% CIs for the majority (8/13) of tests (Fig. 2).

3.2.3. RR verification

RRs derived from the study and verification cohorts were in good agreement for most selected chemistry tests (mean[SD] difference = 0.4% [1.2%]), except Albumin (LRL) and ALP (URL). However for ferritin and TSH, calculated LRLs differed by > 20% (Table 5).

4. Discussion

We have generated method-specific Reference Ranges (RRs) for a comprehensive set of biochemical and endocrinological tests, the majority of which have been implemented into our current service. Herein we report our observations, experience and challenges with this work, as part of laboratory consolidation and verification of new analytical platforms. This RR work has therefore exceeded routine laboratory verification procedures (e.g. method comparison) used for transitioning to new methods. We contend that our overall aim of generating method-specific RRs has been achieved, with the added value of permitting appropriate result interpretation. Our verification process has given useful insight and confidence on the analytical performance of each assay and on our reported RRs.

The outcomes of our approach were not entirely met with parallel support and confidence by our clinical users, in the case of some RRs which seemed unfamiliar compared to RRs used previously. Some RRs caused mis-alignment to thresholds quoted in existing patient management guidelines e.g. local decision thresholds for magnesium (Mg) replacement, reflecting our previously reported lower LRL. Upon subsequent clinical liaison, we explained the effects of inter-method bias on RRs and resultant circumstances where RRs were not transferable between test methods. We also explained how diagnostic thresholds quoted in clinical guidelines and protocols may not account for such discrepancies [14] and therefore inadequately reflect current methodology [15]. However, some published guidelines do acknowledge inter-methodological differences and recommend caution, even if method-specific thresholds are not quoted [16]. For those tests where the RR created uncertainty from our users, these were verified independently, using a data mining approach. The derived RRs closely corroborated those determined in our initial RR study, particularly for chemistry-based tests. It was noteworthy that the confidence intervals (CIs) of the RRs from the verification study were, as expected, much tighter, (literally) giving greater confidence and weight to decisions where both RR approaches showed discernible differences e.g. TSH (LRL) and ALP (URL).

On reviewing the manufacturer's quoted RRs we saw that most chemistry tests (15/23) were not determined in studies involving the manufacturer's own methodology. Instead they tended to refer to other published sources including scientific journals and texts books. For only eight tests did kit inserts refer to RR studies which we only presume have involved Abbott's own methodology. For electrolytes (including calcium), our RRs were corroborated by the verification study as well as Abbott-based studies. For other

Fig. 1. Reference Ranges (RR) for Chemistry tests. For each test, the LRL and URL (—), Confidence Intervals (–) and intervening range (|) are shown. Respective LRLs and URLs quoted by Abbott and Pathology Harmonisation (PH), where available, are also shown for comparison. *Indicates tests where Abbott and/or PH RRs are outside our RR's CIs.

chemistry tests, method specific RRs are justified e.g. LDH, where there is a range of commercially available assays and sub-types e.g. Lactate to Pyruvate and *vice versa* giving very different results (two-fold).

When assays use calibration material that is traceable to certified standards, through a reference measurement system, result comparability across manufacturers' methods may be augmented [10,17]. Such tests are thereby potentially suitable candidates for RR harmonisation. This is the pragmatic approach used by the UK Pathology Harmony Group, for tests showing only minor result differences between methods, assumed to be of little clinical significance [8]. To date, PH has agreed a common set of RRs for a cohort of chemistry based tests. However such harmonisation has not been achieved, nor is valid, for other chemistry tests e.g. ALP due to fundamental assay differences. Nor is it valid for many immunoassay based tests where inter-method non-comparability is well

Table 4

Reference Ranges (RR) for Immunoassay tests. Upper and Lower Reference Limits (URLs and LRLs) quoted by Abbott and those determined in the current study (MMUH – Mater Misericordiae University Hospital) are shown. Female-specific RRs (+ CIs, n = 100–118), as determined using the robust method [7] (where n < 120) are shown in italics. Differences (%) between Abbott and MMUH URLs/LRLs are also shown. [#, ##]Abbott based RRs where the Architect was specifically stated ([#]) or not ([##]). CI = Confidence Interval, NA = Not Available, M = Male.

Test	Abbott		MMUH				MMUH vs Abbott	
	LRL	URL	LRL	90% CI	URL	90% CI	LRL	URL
[##]Anti-Tg (IU/ml)	NA	4.41	0.45	NA	4.8	NA	NA	8.1
[##]Anti-TPO (IU/ml)	NA	5.61	0.0	NA	137.5	NA	NA	95.9
[#]B12 (ng/L)	187	883	194.0	171–250	725.0	679–774	3.6	-21.8
[##]Cortisol (nmol/L)	101	536	120.0	109–153	467.0	438–679	15.8	-14.8
[#]Ferritin (mg/L)								
Female	5	204	*8.0*	*NA*	*193.0*	*166–216*	37.5	-5.7
Male	22	275	29.0	21–41	441.0	403–494	24.1	37.6
[#]Folate (mg/L)	3.1	20.5	3.8	1.1–4.6	18.3	17.7–18.8	18.4	-12.0
[#]FSH (mIU/ml)	0.95	11.95	1.3	0.5–1.6	10.9	9.3–14.2	26.9	-9.6
[##]freeT4 (pmol/L)	9	19	10.5	9.6–11.8	17.9	16.1–23.8	14.3	-6.1
[##]freeT3 (pmol/L)	2.6	5.7	3.6	3.1–3.64	5.4	5.2–5.9	27.8	-5.6
[#]LH (mIU/ml)	1.14	8.75	1.4	1.0–1.53	6.6	5.99–7.54	16.8	-32.0
[##]Prolactin (mIU/L)								
Female	109	557	*114.0*	*NA*	*541.0*	*486–593*	4.4	-3.0
Male	75	407	102.0	91–107	472.0	385–493	26.5	13.8
[#]PTH (intact, pmol/L)	1.59	7.24	1.8	1.4–2.2	7.6	7.4–9.9	11.7	4.7
[##]TSH (mIU/ml)	0.35	4.94	0.21	0.007–0.71	4.9	3.4–5.5	-66.7	-0.8

Difference >20% (red)

Difference 10-20% (amber)

Difference <10% (teal)

known, reflecting inherent variation in method design e.g. antibody specificity, cross reactivity and lack of assay standardisation [10]. The need for method-specific RRs for immunoassay-based tests is therefore not surprising.

PH's pragmatic approach is not an approach likely to be accepted by others [18–20], who observe clinically significant inter-method differences amongst some of the tests included in RR harmonisation projects. The International Measurement Evaluation Programme (IMEP) is a metrological tool which routine laboratories (> 1000 worldwide) have used to analyse Certified Test Samples (CTS) and compare results against reference method assigned values [18]. Considerable result variation from the assigned value has been reported for a number of analytes, including magnesium and albumin (− 20% to + 30% from the target value) [18]. In a
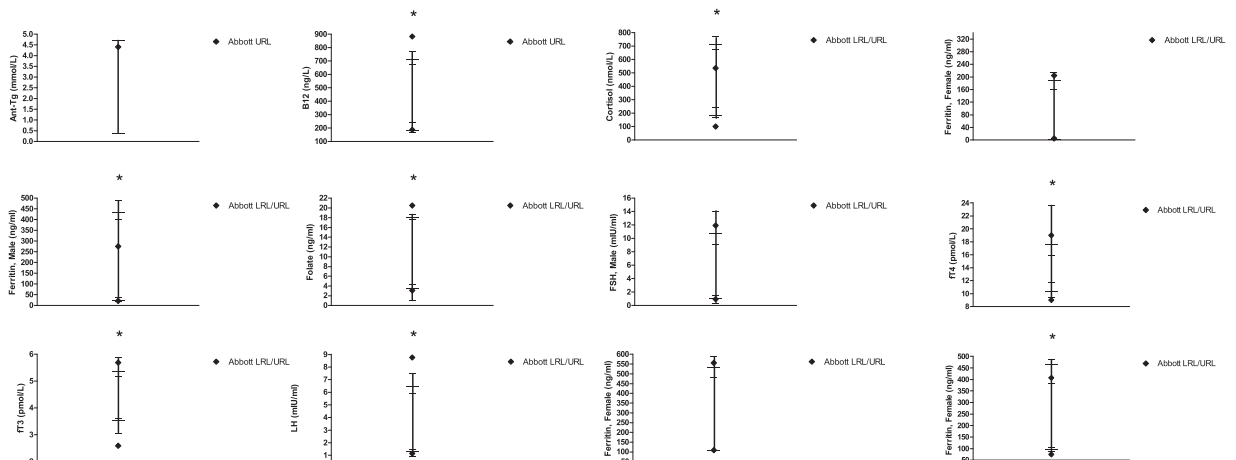

Fig. 2. Reference Ranges (RR) for Immunoassay tests. For each test, the LRL and URL (—), Confidence Intervals (–) and intervening range (|) are shown. Respective LRLs and URLs quoted by Abbott are also shown for comparison. *Indicates tests where Abbott's RRs are outside our RR's CIs. Male FSH and LH RRs are shown.

Table 5

Reference Range (RR) verification. Method-specific Upper and Lower Reference Limits (URLs and LRLs) calculated from the RR Study (RRS) were compared to those determined retrospectively using an independent patient cohort (Verification Study, VS). The same laboratory (MMUH – Mater Misericordiae University Hospital) analytical methodology (Abbott Architect) was used for all analysis. Differences (%) between the URLs/LRLs obtained for the RRS and VS are also shown. CI = Confidence Interval. *Indicates tests where the RR's LRL/URL was outside the CI of the VS's respective limits. #Magnesium and Ferritin (Female) RRs were determined using the Robust method (n < 120) [7].

Test	n	MMUH			
		LRL	90% CI	URL	90% CI
Albumin (BCP, g/L)	4228	32	32-32	44	44-45
Albumin (BCP, g/L)	139	*36	34.3-36.4	44	43.4-44.9
Difference (%)		11.1		0.0	
ALP (U/L)	4210	40	40-41	128	127-131
ALP (U/L)	139	*38	32-46	*108	95-148
Difference (%)		-5.3		-18.5	
Calcium (mmol/L)	2064	2.12	2.11-2.13	2.53	2.52-2.54
Calcium (mmol/L)	134	2.12	2.10-2.20	2.53	2.45-2.58
Difference (%)		0.0		0.0	
Adj. Calcium (mmol/L)	2060	2.19	2.18-2.19	2.55	2.54-2.57
Adj. Calcium (mmol/L)	131	2.18	2.16-2.20	*2.44	2.42-2.46
Difference (%)		-4.6		-4.5	
Magnesium# (mmol/L)	86	0.79	NA	1.02	NA
Magnesium (mmol/L)	138	0.79	0.76-0.80	1.01	0.99-1.06
Difference (%)		0.0		1.0	
Phosphate (mmol/L)	2053	0.78	0.76-0.79	1.43	1.41-1.45
Phosphate (mmol/L)	139	*0.82	0.68-0.87	*1.39	1.35-1.44
Difference (%)		4.9		-2.8	
Ferritin Female (µg/L)	1339	6	6-11	243	219-258
Ferritin# Female (µg/L)	101	8	NA	193	166-216
Difference (%)		20.0		26.0	
Ferritin Male (µg/L)	642	28	23-30	462	416-492
Ferritin Male (µg/L)	128	*29	21-41	441	403-494
Difference (%)		72.4		-4.8	
TSH (mIU/ml)	3105	0.32	0.27-0.36	4.42	4.23-4.55
TSH (mIU/ml)	132	*0.21	0.007-0.71	*4.9	3.4-5.5
Difference (%)		-52.4		9.8	

Difference >20%
Difference 10-20%
Difference <10%

Verification Study (VS)
Reference Range Study (RRS)

subsequent study [19], significant variation in albumin measurements was reported for both bromocresol green- (BCG) and purple- (BCP) based colorimetric methods e.g. Differences of − 3.7 g/L to + 3.2 g/L (designated value = 41.2 g/L) were reported for BCG and up to 4.7 g/L from the designated value (37.4 g/L) for BCP. Considering the clinical use of albumin in adverse risk prediction [20], the authors concluded that such methodological differences render clinical guidelines for interpreting serum albumin applicable only to the albumin method used to develop such guidance [19]. They further advocated for laboratories and manufacturers to generate albumin RRs specific to each method and instrument. We support this proposal, but note (with alarm) that the same RR is quoted for both the Abbott Architect BCP and BCG methods!, Furthermore, PH does not define the albumin method to which their quoted RR pertains, as they do for other chemistry tests showing discernible assay differences e.g. ALP. In our study, the RR for BCP was lower (by 2 g/L at the LRL) than the BCG method, with significantly different method means, agreeing with observations reported elsewhere [19]. Such method differences have other clinical implications, including nutritional assessments, in which scores based on BCG based albumin results indicate better nutritional status than BCP-based scores [21]. Such clinical impact further

justifies the use of method-specific RRs for albumin, rather than a harmonised approach. We are, however, more supportive of other activities and outcomes of PH such as the harmonisation of reporting units [8].

When our method-specific RRs were compared to those of Abbott and PH, a quarter of all chemistry based tests showed discernible differences (> 20%) at the URL (e.g. ALP) and/or the LRL (e.g. iron). However, whilst our verification study also supported a lower URL, Abbott's ALP RR refers to a (non-Abbott) study involving male and female subjects aged > 15 or 20 years of age, respectively. We are unaware of the age of the oldest subjects. In our study, the youngest and oldest volunteers were 28 and 81 years of age, respectively. Variation in such patient demographics may explain reported discrepancies for ALP but are perhaps not the sole cause, given the myriad of factors which affect ALP e.g. body weight, height and cigarette smoking [22], and are unknown to us (in both studies). Furthermore, even when manufacturers employ the same underlying assay principle e.g. International Federation of Clinical Chemistry (IFCC) candidate p-nitrophenol phosphate assay (p-NPP) with addition of aminomethyl propanol (AMP), as specified by PH, significant variation in ALP measurement remains. For a recent distribution (989), from the United Kingdom National External Quality Assurance Service (UKNEQAS), mean ALP concentrations amongst different manufacturers using IFCC methodology (AMP/p-NPP) varied by − 12% to + 24% (119–167 mmol/L) Although non-commutability could have been an issue for this particular distribution, similar intra-method variances for ALP (− 11% to 22%) are also seen for other distributions (e.g. distribution 1007).

For sodium, potassium and chloride, we noted closer agreement between our RRs and those quoted by Abbott. This is perhaps unsurprising since such RRs were derived from Abbott based studies. For all three analytes, PHRRs were wider than those reported by ourselves (and Abbott). Given the volume of testing by diagnostic laboratories for such analytes there could be considerable reclassification of patients depending on the RR used e.g. 20% vs 6.3% incidence of hypokalaemia using a LRL of 3.9 or 3.5 mmol/L, respectively; McGing 2012, personal communication. The extensive Nordic Reference Interval Project (NORIP) [23] quote RRs for sodium and potassium of 137–145 and 3.6–4.6 mmol/L, respectively, which again agree more closely with our method-specific RRs and span a narrower concentration range than PH RRs (6 vs 13 mmol/L for sodium and 1.1 vs 1.8 mmol/L for potassium). In a recent report by Dodd et al. [24] involving Abbott Architect analysis and subsequent data mining, they similarly showed tighter RRs e.g. 137–143 mmol/L for sodium (patients aged 17 years). In their study, the authors commented upon one case involving a child (> 10 years of age) presenting with Addisonian crisis who was considered to have only "mild" hyponatraemia (132 mmol/L). When this figure was evaluated against the PH RR (133–146 mmol/L), suspicion of adrenal insufficiency was reduced. This incident prompted authors to question the validity of using such RRs in children and the study that ensued. The authors raised an interesting point regarding the use of RRs based on a wider 99.7% (than 95%) reference interval, to improve specificity e.g. for patients who are being screened. Such wider RRs could give false reassurance (reduced sensitivity) to some clinical users regarding patient health, particularly patients from the community.

If a laboratory were to report the narrow electrolyte RRs as we and others [23,24] have determined, is it likely that clinical users would find them unfamiliar and question their validity? Probably! Recent clinical practice guidelines define hyponatraemia as < 135 mmol/L, stratifying it further as "mild" (130–135 mmol/L), "moderate" (125–129 mmol/L) and "profound" (< 125 mmol/L) [25], with symptoms more common in the latter group. According to this guideline, such classification is based on "common sense" and "expert experience", not formal evidence. This is likely to agree with the experience of many clinical users. However, the value of any such *biochemical* classification for guiding patient management is perhaps limited since the guideline's algorithms for further investigation (e.g. urine osmolality), differential diagnosis and treatment are *clinically* based, according to severity and onset of symptoms, not the sodium concentration. We do, however, agree that classification of hyponatraemia using biochemical and/or symptomatic criteria (including serum osmolality, speed/duration of development and volume status) might enable practical implementation of the guideline's recommendations. Clearly neither component can be used in isolation to inform management decisions given the possibility of discordance with each other. On balance, we advocate reporting our method-specific RRs as *guidance* regarding the electrolyte status of *healthy* individuals, together with reflective testing e.g. serum osmolality (hypo- vs non-hypotonic) and advice to consider further investigations and treatment, as the guideline recommends.

For the majority of immunoassay-based tests, our method-specific RRs differed by > 20% compared to those reported in Abbott's own studies. For both studies, subjects or data were excluded after considering other relevant laboratory data; Abbott took an *a priori* approach to initially selecting subjects whereas we excluded subjects retrospectively (*a posteriori*) e.g. subjects with abnormal TSH and/or fT4 were excluded either *before* or *after* fT3 analysis, in Abbott's and our studies, respectively. However, even when essentially the same exclusion criteria (and methodology) are used, RRs can still differ considerably as we have shown for fT3 and anti-TPO. Such RR discordance might also reflect differences amongst a manufacturer's range of analyser models. Abbott's FT3 RR relates to analysis on the Axsym to which the Architect shows a constant bias (intercept = 0.23 pmol/L) and might explain (in part) our higher LRL. In Abbott's RR studies, Architects were mentioned specifically in under half (6/13) of the immunoassay-based tests studied. For Anti-TPO, we adopted a data mining approach using patients from general practice and excluded patients with a documented history of thyroid (or sub-clinical) disease, We obtained a lower URL (9.7 IU/mL), in better accord with Abbott (data not shown).

Where the manufacturer's product literature cites a particular RR, this is often accompanied by a recommendation for a laboratory to establish its own range for the population it serves, to account for possible contributions by geographical, environmental and dietary factors. The last factor may be particularly relevant to RRs for haematinic assays. For example, the folate RR (Abbott Architect) used in the USA (7.0–31.4 ng/mL), reflects widespread folate supplementation and is higher than in countries where this does not occur e.g. Ireland and the UK (3.1–20.5 ng/m). For ferritin RRs, discrepancies caused a more pronounced gender difference (male URL > female by 128%) than that reported by Abbott (male URL > female by 35%). This observation may again reflect population-specific factors, genetic differences, physiological and sub-clinical liver processes which were perhaps more prevalent in our male than female subjects. Other contributors to hyperferritinaemia including alcohol consumption, for either sex, were

unknown, but our reported GGT RR was similar to that cited by Abbott (non-Abbott study) as was the magnitude of the gender difference. Patients with either elevated liver enzymes, CRP or raised Mean Corpuscular Volume were excluded when ferritin RRs were determined. It is noteworthy that few subjects were used in Abbott's cited study (32 males and 60 females). Our 90% confidence intervals (403−494) support the observed disparity and such findings, including the pronounced gender difference, were also independently verified.

Our study has several limitations, partly due to its retrospective design, where the scope of analysis was limited by analyte instability e.g. for bicarbonate. This also precluded determination of an anion gap RR. Where some tests e.g. female sex hormones required essential patient information e.g. menstrual cycle stage, generation of meaningful RRs was also precluded. Due to the study size, RR stratification was also limited to gender and, unfortunately, not age. We were unaware of pre-pre-analytical factors such as therapeutic, dietary, nutrition and lifestyle factors, however the latter variables may better reflect our laboratory's patient population. To derive adult reference RRs, we excluded subjects under18 years of age. However, we were unaware of ethnicity, physiological status (e.g. pregnancy) or co-morbidities to apply any further exclusion criteria. For some tests e.g. CK and ALP, the effect of certain demographic factors (e.g black ethnicity and pregnancy) on RRs are well known [11,22]. Inclusion of such subjects has the potential to confound our derived RRs and limit their applicability to other adult patients.

We further acknowledge that for similar subjects and methodology other variables in the pre-analytical (venepuncture, transit, storage etc.) and analytical phases may contribute to differences in RRs. RR discordance due to analytical variation alone may be possible where different studies have used different reagent and calibrators lots, particularly where lot to lot variability is significant. This was observed previously for IGF-1 where successive lot changes culminated in a marked increase (2-fold) in the number of patient results above respective age-related RRs [26].

The justification for using method-specific RRs in promoting accurate interpretation, classification and appropriate patient management also applies to thresholds used for dynamic function tests. In a previous audit on Synacthen testing [27], most respondents (81%) used the same cortisol thresholds (baseline, increment and post Synacthen) despite cortisol's well known between-method bias differences (e.g. negative bias for Abbott Architect). Using inappropriate thresholds to interpret cortisol results has already been shown to cause patient misclassification [28].

Although we demonstrate varying agreement between our RRs and other sources, and postulate reasons for observed discrepancies, we recognize that any consensus does not wholly verify the appropriateness of any RR. We also appreciate that our method-specific RRs will not remain appropriate indefinitely, particularly in view of analytical variation, as discussed above. Such variation would be minimal in our RR study, and elsewhere, where analysis occurs over a relatively short period of time (5 days), involving only one analyser, reagent and calibrator lot for all tests. Intra- and inter-analyser variability may not be commensurate to that seen in routine practice nor over time. The ISO Requirements for Quality and Competence in Medical Laboratories (ISO15189:2012; standard 5.5.5) [12] state that "biological reference intervals shall be periodically reviewed and investigations should be undertaken if the laboratory has reason to believe that a reference interval is no longer appropriate for the reference population"; as might be true for changes in pre-examination or examination procedures. Even without known reasons, and despite the existence of EQA and IQC procedures, we contend that the appropriateness of RRs may still merit periodic consideration or verification as part of a laboratory's commitment to continuous quality improvement so that error can be reduced and maintained at a clinically acceptable level. Our laboratory currently participates in the Empower project [29], which includes an online tool (the "flagger") to enable laboratories to monitor the percentage of results outside RRs or decision limits for any of the 20 tests currently evaluated. Such data could provide laboratories with the necessary prompt to verify existing RRs and this is the focus of our current work.

Declarations

Acknowledgements

References

[1] F. Ceriotti, R. Hinzmann, M. Panteghini, Reference intervals: the way forward, Ann. Clin. Biochem. 49 (2009) 8–17.
[2] M. Plebani, The detection and prevention of errors in laboratory medicine, Ann. Clin. Biochem. 47 (2010) 101–110.
[3] C.G. Fraser, Making better use of differences in serial laboratory results, Ann. Clin. Biochem. 49 (2012) 1–3.
[4] E.K. Harris, Effects of intra-individual variation on the appropriate use of normal ranges, Clin. Chem. 20 (1974) 1535–1542.
[5] C.G. Fraser. Biological variation: from principles to practice, 6th ed., American Association for Clinical Chemistry Inc, Washington DC, AACC Press, 2001.
[6] G.D. Jones, A. Barker, J. Tate, C.-F. Lim, K. Robertson, The case for common reference intervals, Clin. Biochem. Rev. (2005) (2599-104).
[7] G.L. Horowitz, S. Altaie, J.C. Boyd, F. Ceriotti, U. Garg, P. Horn, et al., 2010. Defining, Establishing and Verifying Reference Intervals in the Clinical Laboratory, Approved Guideline. EP28-A3, 3rd ed., Wayne, PA: Clinical and Laboratory Standards Institute.
[8] J. Berg, The approach to pathology harmony in the UK, Clin. Biochem. Rev. (2012) 3389–3391.
[9] R.H. Christenson, D.M. Bunk, H. Schimmel, J.R. Tate. Point: Put simply, standardization of cardiac troponin I is complicated. Clin. Chem., 2012, 58165-8.
[10] C.M. Sturgeon, Common decision limits – the need for harmonised immunoassays, Clin. Chim. Acta 432 (2014) 122–126.
[11] G.R. Jones, Validating common reference intervals in routine laboratories, Clin. Chim. Acta 432 (2014) 119–121.
[12] International Organization for Standardization, ISO 15189 Medical laboratories – Requirements for Quality and Competence, 3rd ed., The Organization, Geneva, 2012.
[13] E.K. Harris, J.C. Boyd, On dividing reference data into subgroups to produce separate reference ranges, Clin. Chem. 36 (1990) 265–270.
[14.] National Collaborating Centre for Women's and Children's Health (UK), National Institute for Health and Clinical Excellence Guidelines. Fertility: Assessment and Treatment for People with Fertility Problems London: Royal College of Obstetrics and Gynaecology (RCOG) Press, February 2013.
[15] J.W. Funder, R.M. Carey, F. Mantero, M. Hassad Murad, M. Rencke, H. Shibata, et al., The management of primary aldosteronism: case detection, diagnosis, and treatment: an endocrine society clinical practice guideline, J. Clin. Endocrinol. Metab. 101 (2016) 1889–1916, http://dx.doi.org/10.1210/jc.2015-4061.
[16] S.R. Bornstein, B. Allolio, W. Arlt, A. Barthel, A. Don-Wauchope, G.D. Hammer, et al., Diagnosis and treatment of primary adrenal insufficiency, J. Clin. Endocrinol. Metab. 101 (2016) 364–389.
[17] G.M. White, Metrological traceability in clinical biochemistry, Ann. Clin. Biochem. 48 (2011) 393–409.
[18] L. Van Nevel, U. Örnemark, P. Smeyers, C. Harper, P.D.P. Taylor IMEP-17: Trace and minor constituents in human serum Eur 2657 Report to Participants. International Comparability. International Measurement Evaluation Programme, 2003 (Part 1).
[19] S.F. Lo, G.W. Miller, B.T. Doumas, Laboratory performance in albumin and total protein measurement using a commutable specimen, Arch. Pathol. Lab. Med. 137 (2013) 912–920.
[20] F.R. Herrmann, C. Safran, S.E. Levkoff, K.L. Minaker, Serum albumin level on admission as a predictor of death, length of stay, and readmission, Arch. Intern. Med. 152 (1992) 125–130.
[21] T. Ueno, S. Hirayama, M. Ito, E. Nishioka, Y. Fukushima, Satoh, T. et al. Albumin concentration determined by the modified bromocresol purple method is superior to that by the bromocresol green method for assessing nutritional status in malnourished patients with inflammation. Ann. Clin. Biochem., 201; 50, pp. 576–584.
[22] T. Gordon, Factors associated with serum alkaline phosphatase level, Arch. Pathol. Lab. Med. 117 (1993) 187–190.
[23] P. Rustad, P. Felding, L. Franzson, V. Kairisto, A. Lahti, A. Mårtensson, et al., The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties, Scand. J. Clin. Lab. Investig. 64 (2004) 271–284.
[24] A. Dodd, N. El-Farhan, S. Moat, Are commonly used paediatric reference intervals for water and electrolyte balance appropriate for clinical use? Ann. Clin. Biochem. 52 (2015) 44–52.
[25] G. Spasovski, R. Vanholder, B. Allolio, D. Annane, S. Ball, D. Bichet, et al., Clinical practice guideline on diagnosis and treatment of hyponatraemia, Eur. J. Endocrinol. 170 (2014) G1–G47.
[26] A. Algeciras-Schmnich, D.E. Burns, J.C. Boyd, J.C. Bryant, K.A. La Fortune, S.K.G. Grebe, Failure of current laboratory protocols to detect lot-to-lot reagent differences: findings and possible solutions', Clin. Chem. 59 (2013) 1187–1194.
[27] K.K. Chatha, J.G. Middle, U.K. Kilpatrick ES. National UK audit of the short synacthen test, Ann. Clin. Biochem. 47 (2010) 158–164.
[28] N. El-Farhan, A. Pickett, D. Ducroq, C. Bailey, K. Mitchem, N. Morgan, A. Armston, et al., Method specific serum cortisol responses to the adrenocorticotropin test: comparison of gas chromatography mass spectrometry and five automated assays, Clin. Endocrinol. 78 (2013) 673–680.
[29] L.A.C. De Grande, K. Goosens, K. Van Uytfanghe, D. Stockl, L.M. Thienpont, The Empower project- a new way of assessing and monitoring test comparability and stability, Clin. Chem. Lab. Med. 53 (8) (2015) 1197–1204.