IFCC Working Group on Standardization of Thyroid Function Tests (WG-STFT)
Meeting at AACC 2008, Washington DC, Monday July 28 (2:00 - 5:00 pm)

PARTICIPANTS

The meeting attendance list is attached in annex.

REPORT ON THE METHOD COMPARISON STUDY FOR TT4, FT4 AND TSH (see also attached powerpoint presentation)

After a warm welcome of the meeting attendees, I (= Linda Thienpont, Chair of the WG-STFT) gave an overview of the general report on the method comparison (MC) study that had been circulated before to the participants in electronic form. As agreed, the data were shown in an anonymous way, however, from the circulated report, each participant knew the code that applied for his/her own company/mass spectrometric laboratory.

The summary of the report per component always concluded on:
(i) the current status of standardization,
(ii) the quality of performance,
(iii) the need for standardization,
(iv) the basis for standardization and the desirable extent of standardization.

TT4

From the MC study it was concluded that the quality of TT4 measurement in the used samples was good for the majority of the assays. In spite of this, some gradation in specificity and susceptibility of the individual assay to sample-related effects could be observed. For 4 assays, standardization would be desirable. Since a complete reference measurement system (RMS) is available for TT4 (see the Database of the Joint Committee for Traceability in Laboratory Medicine (JCTLM) at http://www.bipm.org), this RMS should be the basis for standardization. The panel of single-donation sera (further referred to as “reference panel”) used in this MC study and assigned with values by a JCTLM isotope dilution-mass spectrometry (ID/MS) reference measurement procedure (RMP) can serve as a reference panel for recalibration.

With regard to the approach of recalibration on the basis of a reference panel, the participants exchanged ideas on practices they use within their company. This led to the advice that ideally a manufacturer should include own pools, assign them with values during the measurement of the reference panel, to use them finally as anchor point for sustained standardization. Another advice was that the relationship immunoassay/ID-MS should not only be used as a basis of recalibration, but also to reinvestigate the calibration process as a whole, e.g. in terms of assignment of the zero calibrator, the number and distribution of the calibration points, the number of replicate measurement per calibration point and the curve-fitting model. Finally, it was concluded that ideally, the success of recalibration should be investigated by analysis of a new reference panel.

FT4

From the MC study, it was concluded that the quality of FT4 measurement in the used samples was good for the majority of the assays, although again with some gradation in specificity and susceptibility of the individual assay to sample-related effects. For some assays, the rather poor batch–to-batch consistency and its impact on the magnitude of the within and total precision was highlighted. Most FT4 assays showed considerably lower results than the candidate international conventional equilibrium dialysis (ED) ID-MS RMP and, in addition, they were discrepant amongst each other. The conclusion was that FT4 assays would greatly benefit from standardization, which should be based on a MC study with the candidate ED ID-MS RMP.

A discussion developed about reasons for the difference between the routine assays and the ED ID-MS RMP: i) historically traced manufacturer calibration; ii) insufficient quality of former reference procedures; iii) different measurands. For what concerns the latter, it is a fact routine and reference measure different things. Routine assays estimate FT4 in serum, the RMP measures T4 in dialysate. However, to my opinion (Chair of the WG-STFT), both have the same measurand. ISO, namely, defines the measurand as the *"quantity intended to be measured*" (see ISO/IEC Guide 99:2007 - International Vocabulary of Metrology (VIM)). In this regard, I make reference to another example of standardization of different measurement principles for the same measurand, i.e. standardization of sodium measurement by the direct ion selective electrode (ISE) technology against flame atomic emission spectrophotometry (FAES). Direct ISEs measure the sodium activity in serum water (mmol/kg plasma water), FAES the sodium concentration in plasma (mmol/L plasma) (see NCCLS document C29-A2). However, the convention is to recalibrate ISE to FAES.

Other discussion points/reflections

- Will the international conventional RMP for FT4 generally be accepted?

I confirm in this regard that I agreed with the IFCC Chair of the Scientific Division to prepare a paper in which the IFCC officially recommends standardization of FT4 measurements with the international conventional RMP. Subsequently, I will nominate the RMP for listing by JCTLM. This will require, however, proof of transferability of the procedure set up by UGent. As known to the group, the laboratory collaborating in this regard is that from the Reference Material Institute for Clinical Chemistry Standards (ReCCS), Kawasaki (Japan) (President: Dr. M. Umemoto). Unfortunately, due to other priorities at the ReCCS, the final experiments for investigating the transferability had to be postponed.

- Should clinical samples be included in the standardization process?

It was concluded that it makes certainly sense to use them for investigating the validity of the different assays in cases of abnormal protein binding capacity, however, that this should not necessarily be part of the standardization process.

TSH

From the MC study, it was concluded that the quality of TSH measurements was excellent for the great majority of the assays (for the samples used in this study). Nevertheless, it was observed that several systems lack sensitivity. From the samples measured, there is no evidence that different assays measure different components. Therefore, it was my personal opinion that standardization of TSH measurements is feasible. The MC study showed that, until a RMP will be available, this could be done on the basis of the median of all assays, which would require recalibration of 5 assays.

In the discussion, an opinion was stated that the current status of standardization was sufficient. This was not supported by the majority (e.g., representative from the academic world; representative from the British Thyroid Association). Moreover, the UK experience shows that patients do not accept the current diversity of results and loose confidence in thyroid function tests.

Other important matters of concern

- Will standardization be hampered by the existence of different TSH forms?

This question arises from evidence given in literature that different assays have different specificities to disease-related TSH forms. Therefore, some manufacturers expressed the wish to investigate this. It was argued, however, that this problem is not related to the current standardization project. If different assays have different specificities to different TSH glycoforms, they cannot be standardized anyway.

- How stable will standardization be when it is done using the all methods' median?

According to this concept, there is no anchor other than the current reference panel, which is limitedly available. Standardization can be lost over time, therefore, the study should be repeated from time to time (for example, after 2/5 years). However, care should be taken to measure a successor reference panel overlapping with the first to keep the traceability.

- What will be the new standard (the panel)? Can manufacturers move away from calibration/traceability to the WHO IRP toward traceability to, so to speak, the IFCC STFT MC study 2008? Manufacturers should investigate whether the FDA and/or other regulatory bodies will approve the proposed standardization model.

Because of the importance of the discussion, a paper is intended for Clin Chem Lab Med (attached): please read and report back.

- What about ProficiencyTesting (PT)/External Quality Assessment (EQA)?

In some PT/EQA schemes recovery studies are done with sera supplemented with WHO IRP material. This might become problematic when standardization on the basis of the all methods' median is done. However, on the basis of the expertise of certain participants, recovery experiments are deemed counterproductive. It is known that this practice, when applied with assays for which hierarchically higher RMPs exist, fail for the purpose of accuracy/trueness assessment. It was, therefore, concluded that PT/EQA organizers should be convinced to create new approaches using native samples/pools instead of spiked samples. A tool to convince them could be publications which emphasize the problems/limitations of current PT/EQA systems.

- Why was the median used/was outlier investigation done?

This has been done now. **Please find a attached an EXCEL file** that contains medians and corrected means for TSH. It is recommended to use the corrected mean for recalibration.

WAY FORWARD

Publish on the MC study.
Use the publication as a first tool to convince involved parties about the need and feasibility of standardization of thyroid function tests. The above discussion should be part of the publication. All participating manufacturers/MS laboratories will be mentioned, however, without disclosure of their results. The manuscript draft will be circulated amongst the WG-members and project participants.

Perform a proof-of-concept study
Manufacturers emphasized that they need convincing evidence that the proposed concept works. They need it, in particular, for in-house justification of restandardization. The costs are not so much the panel, but the complete follow-up that is needed when recalibration would be done. I am grateful to Michael Rottman (Roche), who detailed this process upon a question of mine (see below).
The project was not discussed in detail. Therefore, I addressed it in the mail.

Organize a meeting with manufacturers dedicated to the process of recalibration
In this meeting it will be the intention to go into the technical details of the process of recalibration (see above). The meeting may be scheduled in spring 2009.
Note: maybe started in Fortaleza? See e-mail with the question to supply "dose-response" curves for TSH.

Implementation of standardization
- Standardization should be done by all manufacturers at the same time.
- Standardization needs to be sufficiently communicated to all involved parties because it may have drastic consequences: e.g. at the level of interpretation of laboratory data. Therefore, it should be well prepared and properly introduced to all involved parties (clinical chemist, physicians, endocrinologists, patients etc.). It should also be accompanied by a educational process, so that all involved parties receive their information from a primary source, e.g. via publications in dedicated journals, websites, presentations at dedicated symposia etc. Potential parties who could assist in this implementation task are (inter)national clinical and endocrine societies, national health care education programs, clinicians and patient associations etc. Also, regulatory bodies (PT/EQA organizers) should be informed and urged to assess the accuracy/trueness of performance by standardized assays with adequate sample material (see above).

Note on terminology
I think that the term "harmonization" should not be used. It always should be "standardization". It only must be clarified what the standard is: SI; conventional, ….

Manufacturer’s actions when recalibrating an assay

Change in the internal traceability documentation

Shift to a new "reference system-mean value" (hardly to document with sera mainly in the normal range)

Check if the values in the package insert lower detection limit and upper measuring range are affected (if yes separate issue for registration)

Change in target values: primary standards; working calibrators; control samples

Start of workflow to change external control target values

Control of different target values for controls, when you have old and new standardized lots in the market (extreme situation RILIBÄK in Germany)

Change in package insert; regulatory activities due to changes in package inserts (to be submitted to regulatory authorities)

Recalculation of reference values in package insert (to be submitted to regulatory authorities)

Secondary changes in detailed information (pregnancy, children, clinical cohorts, males, females, etc) for reference ranges, for example, in "Thyroid brochures".

Information of customers due to the change in reference value

Change in lab documentation at customer site and information of their clinicians for change in reference range

Check for feed back from the end user

Check for feed back from the country organizations

Annex

IFCC WG-STFT meeting, Washington, DC, July 28, 2008

List of attendees

Name	Affiliation
Michael Rottmann	Roche Diagnostica
Karen Phinney	NIST
Masuo Inoue	Tosoh corporation
Hisao Tsukamoto	Tosoh corporation
Sachiyuki Hasegawa	Tosoh corporation
Judy Ogden	Tosoh Bioscience, Inc.
Susan Kolarik	Tosoh Bioscience, Inc.
Margherita Banci	Diasorin
Jim Faix	Stanford Univ. Medical School
David Miller	Siemens Medical Diagnostics
Wendy Kivens	Beckman Coulter
Trudi Smith	Beckman Coulter
James Sackrison	Beckman Coulter
Philippe Montagne	Biomérieux
Frank Quinn	Abbott
Greg Miller	Virginia Commonwealth University
Tamio Ieiri	Japan TA, Dokkyo Medical Clinn.
Michael Minihan	Olympus
Alan Rockwood	ARUP
Roland Janzen	Siemens
Lothar Siekmann	IFCC-SD-EXEC
Jerald C Nelson	ATA-LSC
Graham Beastall	BTA

Minutes made by:
Prof. Dr. Linda THIENPONT, Chair of the IFCC WG-STFT
Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, UGent
Harelbekestraat 72, B-9000 GENT, Belgium
Tel. +32 9 264 81 04
e-mail: linda.thienpont@ugent.be

Working Group Standardization of Thyroid Function Tests (WG-STFT)

Method Comparison Study for Serum Thyroid Hormones

AACC 2008

Linda Thienpont
Linda.thienpont@ugent.be

Overview

Standardization
(current status, need for standardization, approach)

Method quality

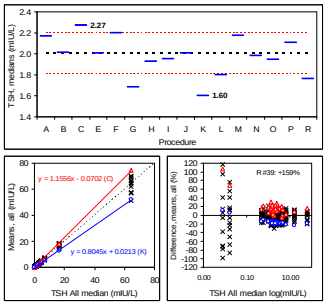
System performance
(imprecision, IQC, stability, batch-to-batch variations)

Miscellaneous
- Correlation FT4/TT4
- Correlation TSH and FT4/TT4

Conclusions

Status of standardization

TSH



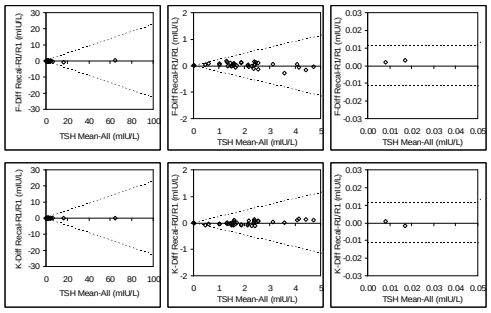
From the 16 methods investigated, 11 gave medians that differed less than 10% from the overall median.

Using the respective regression equations and a reference interval from 0.4 to 4 mIU/L, method K would give TSH values ranging from 0.34 to 3.24 mIU/L, method C from 0.39 to 4.55 mIU/L.

>Standardization is desirable

Do assays measure different components?

Medians: F = 2.2 mIU/L; K = 1.6 mIU/L
Difference plots of recalibrated assays



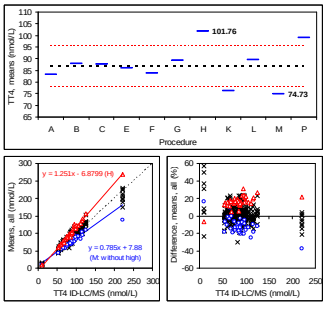
Standardization approach

TSH
- Standardization basis: "All methods median"
- Recalibration required for methods C, G, K, L, R

Future goal
- Standardization within 5%
  - Better batch-to-batch consistency
  - Improvement of IQC practices

Status of standardization

TT4



From the 11 methods investigated, 7 gave means that differed less than 10% from the mean of the reference method.

Means varied from 75 to 102 nmol/L

>Standardization is desirable

Standardization approach

TT4

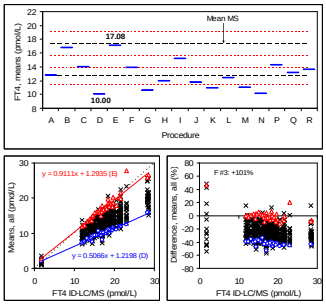
- Standardization basis: Reference method
- Recalibration required for methods H, K, M, P.

Future goal

- Standardization within 5%
  - Better batch-to-batch consistency
  - Improvement of IQC practices

Status of standardization

FT4



Except assays B and E, all assays (15) measured FT4 considerably lower than the MS reference method.

The assays show a great variation of results: means from 10 to 17 pmol/L.

>Standardization is urgently needed

Standardization approach

FT4

- Standardization basis: Reference method
- Recalibration required for nearly all methods

Future goal

- Standardization within 5%
  - Better batch-to-batch consistency
  - Improvement of IQC practices

Method Quality – TSH

Correlation with the all methods median (without the 2 low and the 3 high samples)

Code	r^2	Code	r^2
K	0.998	M	0.991
H	0.997	I	0.990
J	0.997	L	0.990
F	0.993	O	0.988
P	0.993	E	0.987
A	0.992	C	0.961
B	0.992	G	0.955
N	0.992	R	0.948

Best correlations observed for methods K, H and J; worst for methods C, G and R

Method Quality – TSH

Difference plot after recalculation with regression equation

Analytical goals for TSH
(http://www.westgard.com/biodatabase1.htm)

CV (%): 9.7

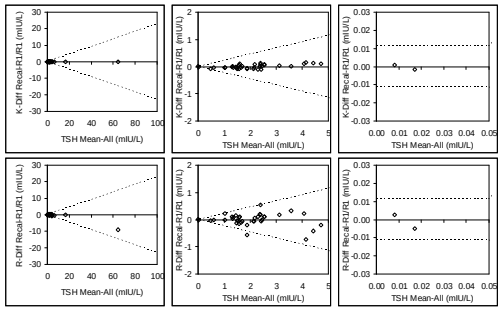
Bias (%): 6.9

Total error (%): 22.8

Extended TE (≤0.05 mIU/L): 0.0114 mIU/L

Method Quality – TSH

Difference plot after recalculation with regression equation: best and worst



Method Quality – TSH

Note

Several assays showed a lack of sensitivity

Method Quality – TT4

Correlation with the reference method (without the low and the high samples)

Code	r^2	Code	r^2
E	0.98	P	0.93
G	0.98	K	0.90
C	0.97	H	0.89
A	0.96	M	0.82
F	0.93	B	0.82
L	0.93		

Best correlations observed for methods E, G, C and A; worst for methods M and B

Method Quality – TT4

Difference plot after recalculation with regression equation

Analytical goals for TT4
(http://www.westgard.com/biodatabase1.htm)

CV (%): 2.5

Bias (%): 3

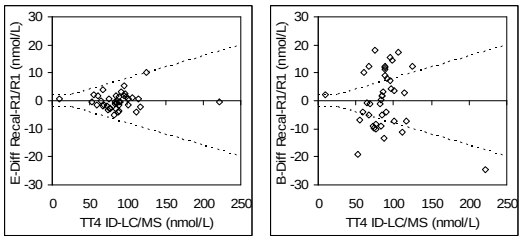
Total error (%): 7.0 (+RM = 8$)

Extended TE (≤25 nmol/L): 2 nmol/L

$Taking the imprecision of the MS method into consideration.

Method Quality – TT4

Difference plot after recalculation with regression equation: best and worst



Method Quality – FT4

Correlation with the reference method (without the low sample)

Code	r^2	Code	r^2
E	0.94	H	0.87
A	0.93	M	0.87
L	0.92	N	0.85
Q	0.92	R	0.85
D	0.90	K	0.84
F	0.89	P	0.83
C	0.88	I	0.82
J	0.88	G	0.74
B	0.87		

Best correlations for E, A, L and Q; worst for G

Method Quality – FT4

Difference plot after recalculation with regression equation

Analytical goals for FT4
(http://www.westgard.com/biodatabase1.htm)

CV (%): 3.8

Bias (%): 3.6

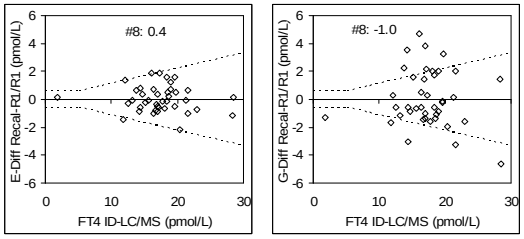
Total error (%): 9.9 (+RM = 11$)

Extended TE (≤5 pmol/L): 0.55 pmol/L

$Taking the imprecision of the MS method into consideration.

Method Quality – FT4

Difference plot after recalculation with regression equation: best and worst



System performance

Imprecision

TSH: The median within run CV ranged from 1.0% (J & K) to 3.1% (G). The median total CV* (1st replicate of each run) ranged from 1.7% (C & N) to 5.4% (F & R).
- **Only from 3 runs, typically with 3 reagent batches but no. of calibrators/instruments variable from MF to MF**

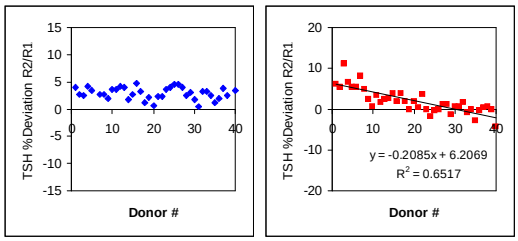
TT4: Median within run CV from 0.8% (E) to 5.7% (H). Median total CV from 1.5% (C) to 6.4% (H).

FT4: Median within run CV from 0.7% (J & P) to 6.7% (G). *Three assays had high within run CV (E: 4.3%, B: 4.9%, and G: 6.7%).* Median total CV from 0.9% (P) to 12.5% (G). *Four assays had high total CV (Q: 6.7%, B: 7.4%, F: 8.0%, and G: 12.5%).*

System performance

Stability

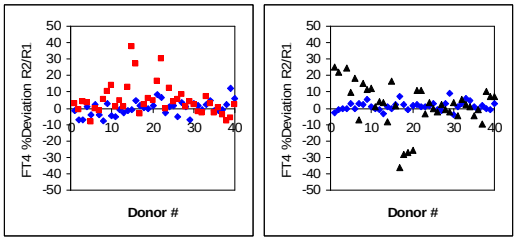
Surprisingly, shifts (left example) and drifts (right example) were seen, in particular, after the measurement of the central IQC samples.



System performance

Batch to batch: increased variation

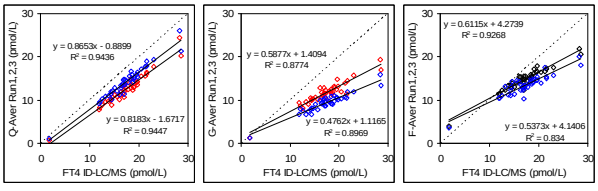
Most problematic for FT4



System performance

Batch to batch: systematic differences between runs

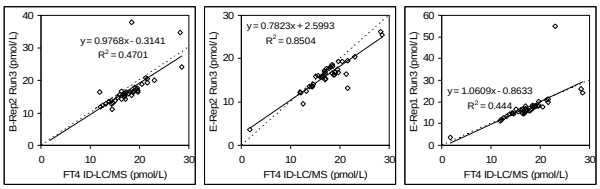
Most problematic for FT4



System performance

Special note FT4

Two assays showed high variation and spuriously high results in some runs



System performance

Internal quality control

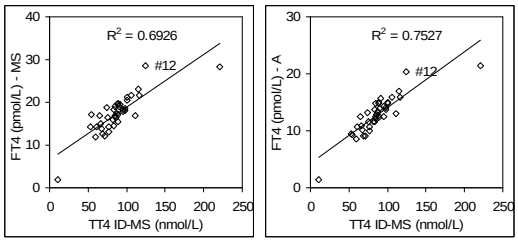
Many were able to control the systems within 5% from the respective targets

Several assays showed considerable deviations from the respective targets

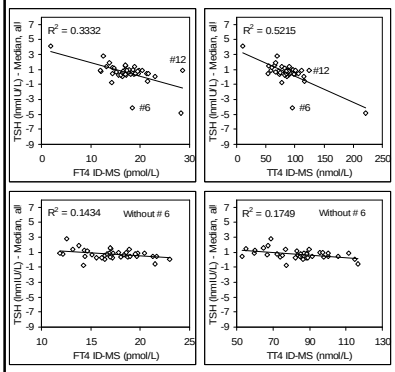
Miscellaneous

Correlation FT4/TT4

The correlation between F- and TT4 is similar in the MS methods and in the routine assays.



Miscellaneous



Correlation TSH and FT4/TT4

Generally, TT4 shows a better correlation with TSH than FT4.

Both show little correlation with TSH in the normal range.

Conclusions – TSH

The quality of TSH measurements is excellent for the great majority of the methods (for the samples used in this study). Nevertheless, several systems lack sensitivity. Standardization (harmonization) of TSH measurements would be beneficial and could be done on the basis of the median of all assays. There is room for improvement of IQC practices and system stability.

Conclusions – TT4

The quality of TT4 measurements is good for the majority of the methods (for the samples used in this study). Standardization of TT4 measurements would be beneficial for 4 methods and could be done on the basis of this study. There is room for improvement of IQC practices and system stability.

Conclusions – FT4

The quality of FT4 measurements was good for the majority of the methods (for the samples used in this study). Standardization of FT4 measurements is necessary and could be done on the basis of this study. There is room for improvement of IQC practices, system stability, and for several methods also for precision.