

Title: Monitoring the Stability of the Standardization Status of FT4 and TSH Assays by Use of Daily Outpatient Medians and Flagging Frequencies

Authors: De Grande, Linde, Kenneth Goossens, Katleen Van Uytfanghe, Barnali Das, Finlay MacKenzie, Maria-Magdalena Patru, and Linda Thienpont

In: *Clinica Chimica Acta* 467: 8–14, 2017

**Title: Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies**

**Authors**: Linde AC De Grande[a], Kenneth Goossens[a], Katleen Van Uytfanghe[b], Barnali Das[c], Finlay MacKenzie[d], Maria-Magdalena Patru[e], Linda M Thienpont[a]* for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT).

[a] Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium.

[b] Ref4*U*, Laboratory of Toxicology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium.

[c] Biochemistry and Immunology Laboratory, Kokilaben Dhirubhai Ambani Hospital and Medical Research Institute, Mumbai, India.

[d] Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK.

[e] Ortho-Clinical Diagnostics, Inc. Rochester, NY, USA.

***Corresponding author**: Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium. Tel.+32-9-264 81 21, Fax +32-9-264 81 98, email: linda.thienpont@ugent.be

**Word count:** 3786

**Figures:** 4; **Tables:** 1

**Abstract (197 words)**

Clinicians diagnose thyroid dysfunction based on TSH and FT4 testing. However, the current lack of comparability between assays limits the optimal use of laboratory data. The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) gave a mandate to the Committee for Standardization of Thyroid Function Tests (C-STFT) to resolve this limitation by standardization. Recently, the Committee members and their partners felt ready to set the step towards the technical recalibration. However, before implementation, they were furthered by the Food and Drugs Administration (FDA) to develop a tool to assess the sustainability of the new calibration basis. C-STFT began to use 2 online applications, i.e., the “Percentiler” and “Flagger”, with the intention to assess their utility for this purpose. The tools monitor the course of instrument-specific moving medians of outpatient results (Percentiler) and flagging rates (Flagger) from data of individual laboratories grouped by instrument/assay peer. They additionally document the mid- to long-term medians, hence, are quality indicators of stability of performance of both laboratories and peers/assays. Here, the first experiences built up in the pre-standardization phase are reported. They suggest the suitability of both applications to document the sustainability of the calibration basis in the post-standardization phase.

**Keywords:** Committee for Standardization of Thyroid Function Tests; Median; Outpatient; International Federation of Clinical Chemistry and Laboratory Medicine; Quality indicator; Population variation.

**Non-standard abbreviations:** C-STFT, Committee for Standardization of Thyroid Function Tests; FDA, Food and Drug Administration; FT4, free thyroxine; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; IVD, in-vitro diagnostic; LIS, Laboratory Information System; TSH, thyroid-stimulating hormone; WG-STFT, Working Group for Standardization of Thyroid Function Tests.

## Introduction

Given the prevalence and severity of different forms of thyroid disease, the yearly number of tests performed worldwide is huge [1-4]. Clinicians mainly rely on the analysis of thyroid-stimulating hormone (TSH) and free thyroxine (FT4) for the diagnosis of thyroid dysfunction and patient follow-up. The frequency of thyroid function testing translates in an enormous impact of the disease on the healthcare system. In this regard, it is generally recognized that, to reduce the expenses for healthcare from laboratory analysis, comparability of measurement data over time, location and across assays would be utmost beneficial. Indeed, once this is achieved, laboratory data can meet modern clinical and public health needs, such as the definition and use of common reference intervals/clinical decision limits, development of evidence-based clinical practice guidelines for application of consistent standards of medical care, translation of research into patient care and disease prevention activities, inclusion of laboratory data in electronic patient records, etc. [5]. However, to accomplish this, in depth transformation of the current laboratory landscape in general but for thyroid function testing in particular is required. Indeed, the problem of observed between-assay discrepancies needs to be resolved [6,7].

The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) decided to pay attention to these needs. In 2005, the Scientific Division formed the Working Group for Standardization of Thyroid Function Tests (WG-STFT) with the mission statement to document the standardization status and intrinsic quality of current thyroid hormone immunoassays. The focus of the activities should be on TSH and FT4 testing, and where necessary, on improving the standardization

status [6,7]. In 2012, the WG was transformed into a Committee (C-STFT) to broaden the scope of stakeholders [8].

The achievements of the WG up to now are described elsewhere [6,7,9-16]. They comprise developing reference measurement systems for standardization of FT4 and harmonization of TSH, demonstrating the feasibility of their use as a uniform calibration basis for commercial in-vitro diagnostic (IVD) thyroid function tests, and designing a step-up approach based on several dedicated method comparison studies to allow new manufacturers to join the standardization activities. Recently the C-STFT set the step to the technical recalibration process of FT4 and TSH assays by a method comparison with clinically relevant panels of samples (results currently under investigation). Although in theory this process completes the establishment of a uniform calibration basis of the assays – at least for diagnosis and follow-up of uncomplicated hypo- and hyperthyroidism – immediate implementation is not possible but needs careful preparation. One of the actions currently undertaken in this regard was that the Committee – comprising laboratory professionals and IVD manufacturers – visited the Food and Drug Administration (FDA) authorities. They presented the technical approach, discussed its acceptability and the plans before implementation. The Committee got a positive response from the FDA, who particularly welcomed the plan to establish a dialogue basis with an as broad spectrum of stakeholders as possible, and investigate with them the benefits but also the risks associated with implementing the standardized/harmonized IVD assays. The benefit-risk analysis recently has been initiated, among others, at the level of medical laboratories (internally by consultation of delegates designated by the IFCC Member Societies) and clinicians/patients [e.g., 17]. In addition, the FDA furthered the Committee to document – preferably at the level of real patient results – the

sustainability of the post-standardization calibration status of the participating IVD assays. The Committee got by courtesy of STT-Consulting and the Chair (currently Thienpont & Stöckl Wissenschaftliches Consulting) access to 2 new quality management tools to assess whether they could serve the above purpose implied by the FDA. Note that as described elsewhere the tools are part of the overarching "Empower project" [18-21]. One tool, called the "Percentiler" monitors daily outpatient medians to reflect the stability/variation of performance at the level of the individual laboratory and its peer group. Its potential to build a global evidence base on IVD test stability across laboratories and peers/manufacturers has been shown before from application for clinical chemistry analytes. The second tool, the "Flagger", monitors flagging of results against reference intervals or decision limits used in the individual laboratory, but also at the peer group level. It is complementary to the "Percentiler" in that it directly translates the effect of analytical quality/(in)stability on flagging as surrogate medical decision making [22]. In view of this potential the C-STFT decided to start using the Percentiler and Flagger (in cooperation with the Empower team) in the framework of its standardization activities. One important matter of concern that needed investigation was whether the tools would similarly be useful to monitor the stability of FT4 and TSH assays as they are for clinical chemistry ones, particularly, because it could be anticipated that the reported median and flagging rate values would be based on a substantially lower number of results per day. The Empower team invited laboratories, already using the applications for clinical chemistry analytes, to extend their participation to FT4 and TSH. For obvious reasons, the focus was on laboratories using the IVD test systems/assays involved in the C-STFT standardization/harmonization project. The C-STFT's intention was to explore the utility of both tools in the pre-standardization phase, and if positive, to fully exploit

them in the post-standardization phase for the purpose implied by the FDA. Here, we report on behalf of C-STFT on the experience built up in the pilot study.

## Material and methods

The way the data are collected in both applications has been described in detail elsewhere [19-21]. Participation is free of charge. In brief, laboratories calculate – preferably by an automatic function in their Laboratory Information System (LIS) or, if not available, manually – instrument-specific daily medians (preferably) from outpatient results. The data are automatically sent by e-mail on a daily basis or batch wise to the Percentiler's and Flagger's MySQL database. For the Flagger application laboratories also report the daily flagging rate in percentage of the total number of generated results. Whereas the Empower team can investigate the complete database at the individual laboratory and peer group level, the participating laboratory only has access to its own data via a user interface (to access via a specific login and password at https://thepercentiler.be and https://theflagger.be, respectively). These interfaces have several functionalities, such as downloadable charts of the laboratory's instrument-specific moving medians of patient results (Percentiler) and flagging rates (Flagger) in time, as well as a table with summary statistics (bias, robust CV). The moving median charts also show the mid- to long-term medians of the laboratory and its peer group. In the Percentiler application the respective numerical values are documented in the statistics table, where also a target median is given (see below). The laboratory bias is compared to the peer group and target median. The deviation (in %) from the target is evaluated against desirable bias limits from biological variation, i.e., 3.3% for serum FT4 and 7.8% for serum TSH, respectively [23]. These limits are visualized in the charts by a gray zone

around the laboratory's mid- to long-term median to reflect the stability of performance. Violations of the limits that last more than 1 week are considered significant. With regard to the aforementioned target median used to assess the bias of the individual laboratory and peer group medians, currently the all-laboratories' median is utilized. In the pre-standardization phase the overall median for FT4 is 15 pmol/L with ± 0.5 pmol/L as limits, for TSH 1.56 mIU/L ± 0.12 mIU/L, respectively. In the Flagger application a certain relative percentage around the long-term median (with an absolute minimum of 1%) is used as the limit, which should not be violated. For FT4 and TSH the limit is preliminary set to 30% [22].

**Results**

In the Percentiler application, currently (March 2016) 76 laboratories participate with 158 instruments, while in the Flagger, 33 laboratories supply data from 44 instruments. The number of laboratories and test systems per manufacturer are listed in Table 1, including the average participation time per peer group. This should give an indication of the number of data points accounted for in this pilot study (1 data point (1 median value) per assay per instrument per day is received). For this study, we distinguished in the Percentiler between 5 peer groups, i.e., Roche Cobas, Siemens Centaur, Abbott Architect, Beckman Synchron, and OCD Vitros, whereas in the Flagger, only the Roche Cobas peer group is currently sufficiently substantiated. Therefore, the data given for this application are very preliminary.

We calculated from the patient data in the Percentiler the respective peer group medians for both FT4 and TSH. For FT4 the peer group medians ranged in the pre-standardization/harmonization phase from ~11.7 to 16 pmol/L, for TSH from ~1.2 to 1.7 mIU/L, respectively. In Figure 1, the match of the peer group medians in this

pilot study with those from the previous Phase I method comparison studies is shown [6,7].

Although the time period of participation is still short for the majority of assays (on average 11 months), most laboratories generally showed a stable performance for both analytes. However, in some individual cases we observed greater variation in performance (drifting or shifting medians), occurrence of a transient bias, between-instrument differences within a laboratory, etc. A few representative examples are given: in Figure 2A a laboratory is documented with an acceptable analytical stability for TSH analysis on all instruments; indeed all moving medians are nicely between the stability limits and concordant with the peer group median; in contrast, Figure 2B shows a laboratory with highly variable FT4 moving medians for all instruments it uses; in Figure 2C the concerned laboratory has clear shifts in its FT4 performance (note the moving medians exceeding the stability limits), while the laboratory shown in Figure 2D performs for TSH with a substantial difference between the 2 instruments it uses.

Figure 3 shows a peculiar observation made in 1 of the peer groups, i.e., 2 subgroups of laboratories having their long-term median at different levels.

When comparing the %-hypo and %-hyper value in the Flagger with the variation of the moving medians of the corresponding laboratories in the Percentiler, the interplaying effect of both tools is visible, i.e., an increase of the median values results in a decrease of the %-hypo and increase of the %-hyper, respectively and vice versa. This is documented in Figure 4, where indeed the upward (until September 2014) and downward trends in the FT4 median values in the Percentiler graph are mirrored in the corresponding %-hypo (5B) and %-hyper (5C) medians of the Flagger.

## Discussion

This pilot study was intended to apply the Percentiler and Flagger – initially developed for clinical chemistry analytes – also for FT4 and TSH. We were particularly interested to learn whether the tools can serve the purpose of monitoring/demonstrating the stability of the assays' calibration status. This kind of tools were indeed furthered from the C-STFT by the FDA as part of a benefit-risk analysis before implementing the recalibrated assays. As previously described, the big advantage of both Percentiler and Flagger is that they work with results from patient samples. This prevents questioning of the observations because of non-commutability issues typically associated with processed materials used in external quality control surveys conducted for the same purpose [24]. That said, there may well be merit in looking at this data in conjunction with data from a mature, frequent distribution, data-rich external quality assurance services program which uses material at the more commutable end of the spectrum and which already regularly produces method trend data [e.g., 25]. It also circumvents discussions whether internal quality control data are sufficiently suited to reflect variation in laboratory/instrument/assay performance due to, for example, reagent and calibrator lot changes [26]. Although we conducted this study in the pre-standardization phase in which all FT4 and TSH assays still work with their original calibration basis, our reasoning was that, if positive, the tools would likewise be useful in the post-standardization phase. On long-term we aim at having all manufacturers/instruments/assays involved in the C-STFT activities represented in the Percentiler and Flagger application by a sufficient number of laboratories (we aim at an input of data from minimum 20 instruments per manufacturer for solid peer

group observations). In this pilot phase we could only distinguish 6 peer groups, of which 5 still are absolutely preliminary, which requires cautious interpretation (data input from too few instruments). Nevertheless, we believe that also observations from an exploratory phase are helpful to build experience. We are confident that the IVD partners of C-STFT will be eager enough to bring more customers on board. After all, meeting the FDA demands might facilitate the revision of the 510k clearance of their recalibrated assays.

We first explored the utility of the Percentiler to do quasi real-time monitoring of FT4 and TSH outpatient medians in the individual laboratory. This monitoring is a quality indicator of stability of performance of both the laboratory and assay. We used the experience from applying the tool in clinical chemistry for comparison [18-20]. As previously explained, the on-line user interface shows the participating laboratory for each instrument the course of the moving median, the mid- to long-term median as well as that from the peer group. Interpretation of the graphs in terms of acceptable performance is facilitated by including a stability zone around the long-term median. The limits of that zone are desirable bias limits inferred from the biological variation model [23]. Our short time experience with the thyroid hormone assays learns that in most laboratories the stability of performance was quite satisfactory (see Figure 2A). Indeed, no significant or only borderline violation of the FT4 and TSH limits was observed in 80% of the laboratories participating for minimum 6 months; for TSH this was the case for >95% of the participants. Nevertheless, we want to emphasize that the variability of the moving median for FT4 and TSH was higher than in comparison to that for clinical chemistry analytes. As explained before, this was anticipated from the fact that the respective daily medians are calculated from fewer results than is the case for the common clinical chemistry analytes, simply because thyroid hormone

measurements are requested less frequently. However, the increased variability due to fewer results can partially be solved by calculating the moving median from a higher number of daily medians (the options are n = 5, 8 and 16). In other cases this is not effective, most probably pointing to a real increase in analytical variability (e.g., Figure 2B). Hence, we suggest that the application of the Percentiler in the post-standardization phase might better serve the purpose, if we could focus on laboratories with a high throughput. In approximately 20%, 8% and 3% of the laboratories participating during ≥ half a year, we observed respectively 1, 2 or 3 significant event(s) violating the FT4 stability limits (note the events observed for TSH in 5% of the laboratories were borderline). Upon investigation by the concerned laboratories, this was either due to lot changes, recalibration, but mostly reagent instability; also differences between instruments were observed (see Figure 2C and 2D). Nevertheless, we want to repeat that the observations still must be interpreted with caution [19,20]. Consultation with the concerned laboratory is necessary to be sure that, for example, the observed discrepancy between instruments is not due to the fact that they are used for preferential measurement of certain patient samples. Indeed, from contacting participants in the Percentiler application for clinical chemistry analytes, we learnt that laboratories sometimes concentrate the samples from, for example, policlinic patients presenting themselves in the morning for measurement on 1 instrument, while they reserve other instruments for measurement of, for example, day clinic patients (sometimes also identified as outpatient by the LIS).

Some may argue that the violations are due to the fact that we use desirable bias limits inferred from biological variation, which are particularly narrow for FT4. However, as discussed above the violation rate was never of an extent that the

validity of the limits needed to be questioned. In the Percentiler, a laboratory's FT4 and TSH median values are compared with those of the peer group to which it belongs. This allows the laboratory to infer whether a shift or drift is due to its own performance (event only seen for the concerned laboratory), or rather to an assay/manufacturer event (e.g., a reagent or calibrator lot change applying for several laboratories of the peer group). However, to compare the participant with its peer, there are 2 prerequisites. First the peer group median should be sufficiently solid, which depends on the number of instruments used to calculate it (we aim at a minimum of 20). As mentioned before, our pilot study faced in this regard a problem, which prevents us to discuss here the performance of individual laboratories in comparison to their peer in greater detail. Nevertheless, we refer to the example in Figure 2A showing a laboratory performing in concordance with the peer group for nearly 1 year. The second prerequisite is that the medians should be calculated from outpatient results as discussed before [19,20]. Also the sample type analyzed in the participating laboratories may impact the medians. Currently we do not distinguish between medians from serum and plasma sample analysis, because only few participants measure plasma. However, if in the future that number would increase, it might be necessary to make sample-specific subgroups, to prevent that the differences in medians are interpreted as a calibration bias. The prerequisite of medians from outpatient results might apply even more for FT4 and TSH than for clinical chemistry analytes, because disease- or patient population-related concentrations might be quite influential. This is nicely illustrated from the above reported observation of 2 subgroups for TSH within the same peer group (Figure 3). In our opinion there were 2 explanations possible: either it was due to a real instrument bias in the subgroups, or to the way outpatients are defined in the

subgroup laboratories. To discriminate between these bias sources, we let analyze a same set of 20 samples in a laboratory from either subgroup. No significant bias between those 2 laboratories was found, from which we concluded that the observed difference was most likely due to a different patient population served by the 2 subgroups, i.e., 1 group of laboratories measures TSH mostly for screening purposes, while the other does the measurements rather for follow-up of therapy. Some probably will see this as a limitation of the utility of the Percentiler to assess the bias of peer groups in the post-standardization phase. For many laboratories it is currently still difficult to unequivocally define results from outpatients due to limitations of their LIS. However, we are confident that it will be possible to resolve this weakness in the future, as we found already several LIS providers willing to adequately adapt their data transfer logic. Market forces most probably will make the others to follow.

In second instance, we explored the utility of the Percentiler to reflect the calibration status of IVD test systems/assays with particular emphasis on the sustainability thereof and/or the comparability between manufacturers. As further illustrated in Figure 1, it was comforting to observe how remarkably the medians of the 6 peer groups considered in this pilot study matched those seen in our Phase I studies [6,7]. The medians of these previous studies were perfectly suited for comparison, because they were calculated from results of method comparison studies with samples from apparently healthy (euthyroid) volunteers. Note that here we again report anonymously on the peer group data – as we did in all previous reports [6,7,14,16] – simply because the current study was only exploratory and did not yet include all manufacturers/instruments test systems. Anyhow, we see the observations in Figure 2 as a first basis of evidence for the utility of the Percentiler to

serve the purpose implied by the FDA. We expect this evidence to increase, the more solid the peer groups become (lower variability of the medians) [19,20]. Once we will have a sufficient number of participants for the different peer groups, we will be able to monitor their stability. Significant events observed for 1 or several of the instruments/assays will be used as an indication that the standardization/harmonization status might be jeopardized. Alternatively, it might point to a too high lot-to-lot variability. Therefore, these observations should form the basis for in-depth discussions on the lot to lot variability with the concerned manufacturers or be an incentive to conduct a new method comparison study to realign the shifted calibration basis. For this purpose, the C-STFT already prepared a follow up panel for TSH and works on an additional follow up panel for FT4.

In the statistical table in the user interface, the individual laboratory- and peer group bias are assessed against the all-laboratories median targets for FT4 and TSH. This is a logical target in the pre-standardization phase, however, in the post-standardization phase it will be adapted to the FT4 targets assigned by the conventional reference measurement procedure and the TSH all-procedure trimmed mean inferred from the method comparison on the harmonization panel.

Although our experience with the Flagger is still preliminary (only few laboratories send data on the flagging rate), we presume that the tool will be useful to investigate the impact of analytical quality/instability on daily surrogate medical decision making in the post-standardization phase, exactly as it does for clinical chemistry analytes [21]. This is, for example, already now nicely demonstrated for the case shown in Figure 4, where the fluctuations in the patient median values are mirrored in the flagging rate medians. The Flagger application is not strictly needed to assess the sustainability of the standardization status, but, we want to offer the

Percentiler-Flagger link as an interesting option to the participating laboratories. It indeed allows a laboratory to react rapidly on observed changes in the flagging rate, even if the underlying analytical instability is not yet considered significant [21]. Laboratories appreciate that they can prevent complaints from clinicians about an abnormal increase of the number of flagged results. But even if there are complaints, the Flagger might serve the laboratory to document that the perception of the clinician is not correct. We currently use a stability limit of 30% around the long-term median in the Flagger, but fine-tuning will be done after a longer follow-up.

## Conclusion

By starting to use the Percentiler and Flagger application in the framework of the C-STFT activities, their utility for monitoring the calibration basis of FT4 and TSH assays in the pre-standardization status is shown. This looks promising for their utility in the post-standardization phase to monitor the sustainability of the recalibration status achieved through the C-STFT project. Nonetheless, the here described limitations need to be resolved, mainly by expansion of the number of participating laboratories preferably with a high throughput, representation of all manufacturers on the project, and better definition of outpatient results.

## Acknowledgments

The authors are extremely indebted to the laboratories that are participating in this study. They also express their gratitude to Thienpont & Stöckl Wissenschaftliches Consulting for receiving access to the Percentiler and Flagger database.

## Declaration of interest

Finlay MacKenzie is Organiser of the UK NEQAS for Thyroid Hormones.

## References

[1] Golden SH, Robinson KA, Saldanha I, Anton B, Ladenson PW. Clinical review: Prevalence and incidence of endocrine and metabolic disorders in the United States: a comprehensive review. J Clin Endocrinol Metab 2009;94:1853-78.

[2] Vanderpump MP. The epidemiology of thyroid disease. Br Med Bull 2011;99:39-51.

[3] Bjoro T, Holmen J, Krüger O, Midthjell K, Hunstad K, Schreiner T et al. Prevalence of thyroid disease, thyroid dysfunction and thyroid peroxidase antibodies in a large, unselected population. The Health Study of Nord-Trondelag (HUNT). Eur J Endocrinol 2000;143:639-47.

[4] Leese GP, Flynn RV, Jung RT, Macdonald TM, Murphy MJ, Morris AD. Increasing prevalence and incidence of thyroid disease in Tayside, Scotland: the Thyroid Epidemiology Audit and Research Study (TEARS). Clin Endocrinol 2008;68:311-6.

[5] Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067-75.

[6] 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.

[7] 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.

[8] IFCC Scientific Division, SD-Committees. Standardization of Thyroid Function Tests (C-STFT). http://www.ifcc.org/ifcc-scientific-division/sd-committees/c-stft/ (accessed November 2015).

[9] Thienpont LM, Beastall G, Christofides ND, Faix JD, Ieiri T, Miller WG et al. International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), Scientific Division Working Group for Standardization of Thyroid Function Tests (WG-STFT). Measurement of free thyroxine in laboratory medicine - proposal of measurand definition. Clin Chem Lab Med 2007;45:563-4.

[10] Thienpont LM, Beastall G, Christofides ND, Faix JD, Ieiri T, Jarrige V et al. Proposal of a candidate international conventional reference measurement procedure for free thyroxine in serum. Clin Chem Lab Med 2007;45:934-6.

[11] Van Uytfanghe K, Stöckl D, Ross HA, Thienpont LM. Use of frozen sera for FT4 standardization: investigation by equilibrium dialysis combined with isotope dilution-mass spectrometry and immunoassay. Clin Chem 2006;52:1817-21.

[12] International Federation of Clinical Chemistry; Laboratory Medicine Working Group for Standardization of Thyroid Function Tests, Van Houcke SK, Van Uytfanghe K, Shimizu E, Tani W, Umemoto M, Thienpont LM. IFCC international conventional reference procedure for the measurement of free thyroxine in serum: International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Working Group for Standardization of Thyroid Function Tests (WG-STFT)(1). Clin Chem Lab Med 2011;49:1275-81.

[13] 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.

[14] Thienpont LM, Van Uytfanghe K, Van Houcke S; IFCC Working Group for Standardization of Thyroid Function Tests (WG-STFT). Standardization activities in the field of thyroid function tests: a status report. Clin Chem Lab Med 2010;48:1577-83.

[15] Van Uytfanghe K, De Grande LA, Thienpont LM. A “Step-Up” approach for harmonization. Clin Chim Acta 2014;432:62-7.

[16] 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.

[17] Thienpont L, Faix J, Beastall G. Standardization of free thyroxine and harmonization of thyrotropin measurements: A request for input from endocrinologists and other physicians. Thyroid 2015;25:1379-80.

[18] Van Houcke SK, Stepman HC, Thienpont LM, Fiers T, Stove V, Couck P et al. Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.

[19] De Grande LA, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont LM. The Empower project - a new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015;53:1197-204.

[20] Goossens K, Van Uytfanghe K, Twomey PJ, Thienpont LM; Participating Laboratories. Monitoring laboratory data across manufacturers and laboratories - A prerequisite to make "Big Data" work. Clin Chim Acta 2015;445:12-8.

[21] Goossens K, Brinkmann T, Thienpont LM. On-line flagging monitoring - a new quality management tool for the analytical phase. Clin Chem Lab Med 2015;53:e269-70.

[22] Stepman HCM, Stöckl D, Twomey PJ, Thienpont LM. A fresh look at analytical performance specifications from biological variation. Clin Chim Acta 2013;421:191-2.

[23] Westgard QC. Biological variation database, and quality specifications for imprecision, bias and total error (desirable and minimum). The 2014 update. http://www.westgard.com/biodatabase-2014-update.htm (accessed November 2015).

[24] Miller WG, Myers GL. Commutability still matters. Clin Chem 2013;59:1291-3.

[25] UK NEQAS – International Quality Expertise. Thyroid hormones. http://www.uknegas.org.uk/content/Pageserver.asp (accessed November 2015)

[26] Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. Clin Chem 2011;57:76-83.

**Table 1:** Overview of the number of instruments and the average participation time (in months) per peer group/manufacturer in the Percentiler and the Flagger.

| | The Percentiler | | | The Flagger | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| | **Participants** | **Instruments** | **Average participation time (months)** | **Participants** | **Instruments** | **Average participation time (months)** |
| | | | | | | |
| **Total** | 76 | 158 | 11 | 33 | 44 | 6 |
| | | | | | | |
| **Abbott Architect** | 10 | 19 | 10 | 2 | 2 | 6 |
| **Beckman Synchron** | 11 | 15 | 11 | 8 | 9 | 6 |
| **OCD Vitros** | 5 | 11 | 8 | 2 | 2 | 4 |
| **Roche Cobas ElecSys** | 38 | 81 | 11 | 15 | 22 | 5 |
| **Siemens Centaur** | 8 | 25 | 11 | 3 | 4 | 4 |
| **Siemens Vista** | 4 | 7 | 9 | 3 | 5 | 7 |

## Legend to Figures

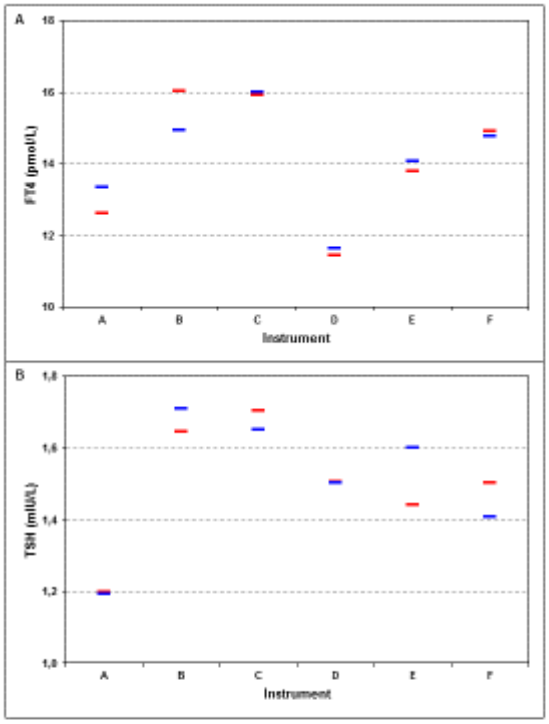


**Figure 1** Illustration of the match between the median values per manufacturer/instrument for FT4 and TSH in the Percentiler application in this study (blue lines) and those in the Phase I method comparison study (red lines) [6,7].

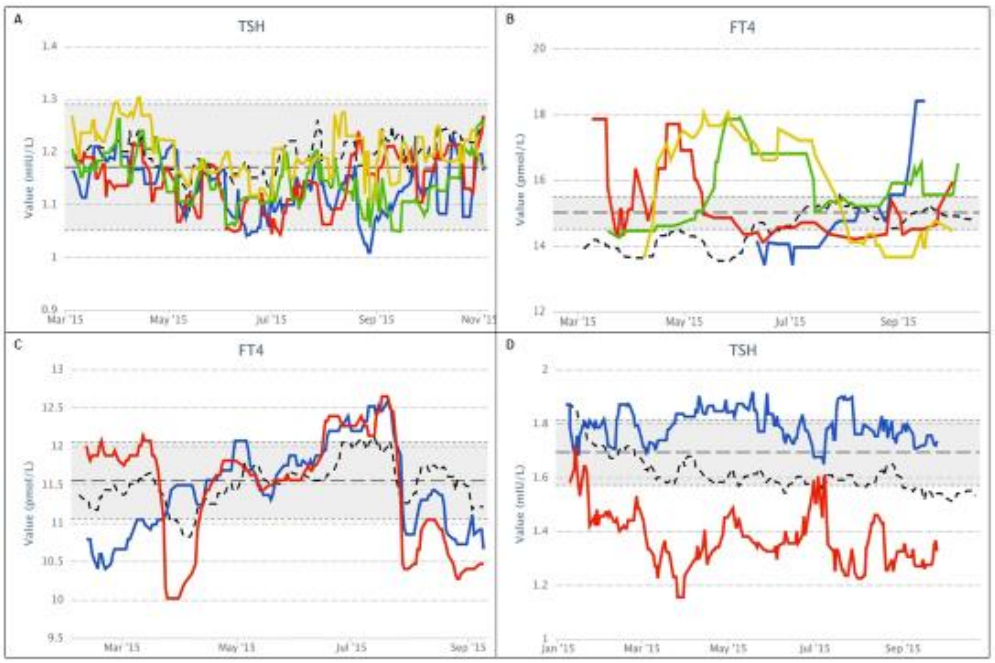


**Figure 2** Examples of Percentiler graphs. Each colored line represents a single instrument in a laboratory; the long broken gray line shows the long-term median of the laboratory, while the black short broken line represents the peer group moving median. The shaded area between the short broken gray lines is the so-called stability zone; violation for longer than one week is considered significant. In (A), we show the time course of the TSH moving medians for all instruments used in a certain laboratory; for all instruments the analytical variability is low, nicely between the stability limits, and the medians are, in addition, concordant with the peer group median for nearly a year; in (B) we demonstrate a laboratory with a highly variable FT4 performance for all instruments; in (C) we document a laboratory with clear shifts in the FT4 moving medians outside the stability limits; in (D) we demonstrate that

sometimes a laboratory performs with a substantial difference in the medians of the instruments it uses (here for TSH) (is to interpret with caution, though as explained in the discussion).

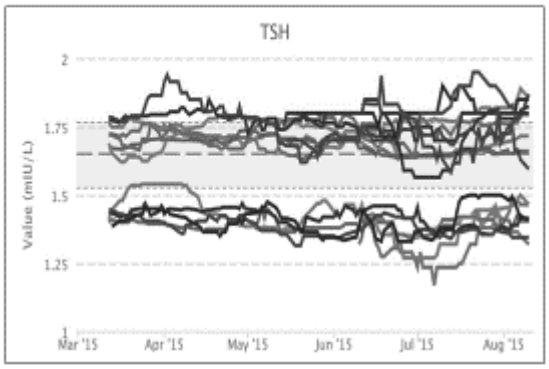


**Figure 3** Occurrence of two subgroups in a single TSH peer group. The graph shows the time course of the TSH moving medians for all instruments from several laboratories in a certain peer group. Each line represents a single instrument; the long broken gray line shows the long-term median of the entire peer group. The shaded area between the short broken gray lines represents the stability zone of ± 0.12 mIU/L around the peer group “overall” median. The graphs clearly show the occurrence of two subgroups within a single peer group, one having the long-term median around 1.4 mIU/L, the other around 1.75 mIU/L.

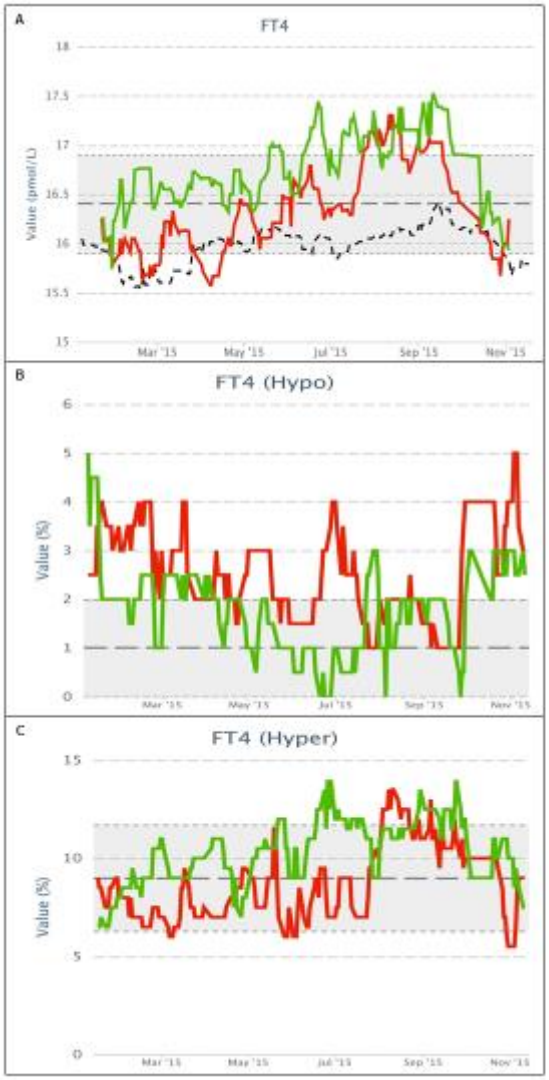


**Figure 4** Percentiler graph showing an upwards drift for the moving median of FT4 for two instruments (A), which is mirrored in the Flagger by a decreasing %-hypo (B) and increasing %-hyper value (C). When the moving median decreases again around September (A), the %-hypo increases (B) and the %-hyper decreases (C). The long broken black line (A) shows the peer group median, while the long broken gray line shows the long-term median of the patient medians (A) and flagging rates (B and C). The shaded area between the short broken gray lines represents the stability zone, which should not be violated longer than one week.