How low can we (reliably) go? A method comparison of TSH assays with a focus on low concentrations.

Stan R. Ursem¹,², Anita Boelen¹,²,³, Jacquelien J. Hillebrand¹,², Wendy P.J. den Elzen⁴,⁵, Annemieke C. Heijboer¹,²,³,⁶

¹ Amsterdam UMC location University of Amsterdam, Department of Laboratory Medicine, Endocrine Laboratory, Meibergdreef 9, Amsterdam, The Netherlands
² Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam, The Netherlands
³ Amsterdam Reproduction & Development, Amsterdam, The Netherlands
⁴ Amsterdam UMC location University of Amsterdam, Department of Laboratory Medicine, Laboratory Specialized Diagnostics & Research, Meibergdreef 9, Amsterdam, The Netherlands
⁵ Amsterdam Public Health, Amsterdam, The Netherlands
⁶ Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Laboratory Medicine, Endocrine Laboratory, Boelelaan 1117, Amsterdam, The Netherlands

Corresponding author:
Prof. Dr. Annemieke C Heijboer
E-mail address: a.heijboer@amsterdamumc.nl
Postal address: Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

**Word count:** 1669

**Short title:** TSH assay (low range) method comparison

**Keywords:** TSH, thyroid-stimulating hormone, method comparison, subclinical hyperthyroidism

**List of Abbreviations:** TSH, thyroid-stimulating hormone; CV, coefficient of variation; WHO, World Health Organization
Abstract

Objective
International guidelines concerning subclinical hyperthyroidism and thyroid cancer advice absolute cut-off values for aiding clinical decisions in the low range of thyroid-stimulating hormone (TSH) concentrations. As TSH assays are known to be poorly standardized in the normal to high range, we performed a TSH assay method comparison focusing on the low range.

Methods
Sixty samples, selected to cover a wide range of TSH concentrations (<0.01 to 120 mIU/L) with oversampling in the lower range (<0.4 mIU/L), were used for the method comparison between three TSH immunoassays (Cobas, Alinity and Atellica). In addition, twenty samples were used to assess the coefficient of variation from duplicate measurements in these three methods.

Results
The TSH immunoassays showed standardization differences with a bias of 7-16% for the total range and 1-14% for the low range. This could lead to a different classification of 1.5% of all measured TSH concentrations <0.40 mIU/L measured in our laboratory over the last six months, regarding the clinically important cut-off value of TSH=0.1 mIU/L. As the imprecision of the immunoassays varied from 1.6-5.5% this could lead to a similar reclassification as the bias between immunoassays.

Conclusions
We established the standardization differences of frequently used TSH assays for the total and low concentration range. Based on the proportional bias and the imprecision, this effect seems to have limited clinical consequences for the low TSH concentration range. Nevertheless, as guidelines mention absolute TSH values to guide clinical decision making, caution must be applied when interpreting values close to these cut-offs.
1. Introduction

Thyroid disorders are common. When a thyroid disorder is suspected, the standard procedure is measuring the serum or plasma thyroid-stimulating hormone (TSH). Accurate TSH measurements are important for timely diagnosis since symptoms can be subtle, and also for monitoring therapy in hypoor hyperthyroidism and assessing suppressive therapy in patients with thyroid cancer. Unfortunately, TSH assays are poorly standardized to date.

The IFCC Working Group on Standardization of Thyroid Function Tests established in 2010 standardization differences up to 39% between TSH assays and recommended further research into the concentration range close to the limit of quantification. The working group deemed availability of TSH reference material in the mid- to short term technically unlikely and now focuses on statistical harmonization. Currently, the grand majority of TSH assays are not yet harmonized, which means that absolute concentrations reported by laboratories using different assays are not comparable. This could impact diagnosis, prognosis and management of disease, as has recently been shown for subclinical hypothyroidism.

The lower range of the TSH assay is important, because in this pathophysiological range often clinical decisions are made, and this is also where assay performance decreases. Several international guidelines mention absolute TSH concentrations for clinical decision making such as in subclinical hyperthyroidism, thyroid cancer, pregnancy and assisted reproduction. However, to our best knowledge, research is lacking concerning the comparability of results in this low TSH range.

Therefore, we performed a method comparison for commonly used TSH immunoassays, with oversampling in the lower concentration range (<0.4 mIU/L) and established the imprecision in this low concentration range. Based on the standardization differences, we investigated the theoretical impact for the absolute cut-offs mentioned in guidelines.
2. Material and methods

2.1 Serum samples
Leftover samples were collected from routine measurements in our tertiary center (Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands). Samples were selected to cover a wide range of TSH concentrations, with oversampling in the lower range, based on measurements performed using the Cobas® assay (Roche Diagnostics, Mannheim, Germany), and were anonymized. Lithium heparin plasma samples were used for analysis. Aliquots were stored at -20°C and measured for the current study within two months. Sixty samples were used for the method comparison and twenty samples were used for assessing the coefficient of variation.

2.2 Immunoassays
TSH was measured in one run using the automated electrochemiluminescence immunoassays of the following platforms: Cobas® (Roche Diagnostics, Mannheim, Germany) Elecsys TSH REF 08429324 190, Alinity® (Abbott Laboratories, Abbott Park, IL, USA) TSH reagent kit G71292R02 and Atellica® (Siemens Healthcare Diagnostics, Erlangen, Germany) TSH3-UL REF 10995703. The Cobas and Alinity assays are standardized against the same WHO 2nd International Standard for human TSH (IRP 80/558) and the Atellica assay is standardized against the WHO 3rd International Standard for human TSH (IRP 81/565). The samples were included based on the initial measurement using the Cobas®. For the method comparison all samples were remeasured on the Cobas® platform in one run.

2.3 Statistical analysis
Statistical analysis was performed using Passing-Bablok regression to minimize the effect of outliers possibly caused by assay-specific interference. In addition, Spearman’s rank correlation coefficient was used. Coefficient of variation from duplicate measurements was used to assess intra-assay variation using the within-subject standard deviation method. Statistical analysis was performed using MedCalc® version 18.5. p-Values <0.05 were considered to reflect statistical significance. In addition, we estimated the effect of the greatest standardization difference on measured values around the absolute cut-off of TSH=0.1 mIU/L using the slope and intercept of the Passing-Bablok regressions for the low concentrations. We calculated the percentage that would cross the absolute cut-off, when it would theoretically be measured by another assay. We used all reported TSH results in our academic hospital in six months (1 March 2022 until 1 September 2022).
3. Results

The distribution of the samples, as measured using the Cobas assay are depicted in Supplementary Figure 1. As the lower concentrations were oversampled, the median TSH concentration used for the method comparison was 0.24 mIU/L (interquartile range: 0.04–3.21 mIU/L; n=60). For the duplicate measurements the median was 0.25 mIU/L (interquartile range: 0.06–2.66 mIU/L; n=20).

Passing & Bablok regression and correlation coefficients between the three methods are depicted in Figure 1. The slope of the Passing-Bablok regressions indicate the standardization differences, as the y-intercept was negligible. The bias ranged from 7% to 16% in the complete measured range. For the samples with TSH <0.40 mIU/L, the bias differed between 1% and 14%, as is depicted in Figure 1.

The imprecision of the assays increased as the concentration lowered and is shown in Table 1. Differences were seen between assays and concentrations, with the largest mean variation between duplicate measurements being 5.5% for the Atellica assay in the range [TSH] <0.40 mIU/L.

The Atellica and Cobas assay exhibited the largest bias in the lower concentration range (14%). Based on this bias, 1.5% of all TSH measurements with results <0.4 mIU/L (15 of the 998 measurements) in our laboratory, using the Cobas assay, would cross the absolute cut-off and would be <0.10 mIU/L instead of ≥0.10 mIU/L, when measured with the Atellica assay.

In addition, we compared the initial TSH measurement on the Cobas, which we used to include the samples, to the remeasurements performed on the Cobas, but in one run. The Passing and Bablok regression was $\text{TSH}_{\text{Cobas remeasurement}} = 0.935 \times \text{TSH}_{\text{Cobas initial}} + 0.0003; \rho = 0.998; P < 0.0001$ for all concentrations, and for TSH <0.4 mIU/L: $\text{TSH}_{\text{Cobas remeasurement}} = 0.924 \times \text{TSH}_{\text{Cobas initial}} + 0.0013; \rho = 0.991; P < 0.0001$. This remeasurement would also lead to 1.5% of all TSH measurements <0.4 mIU/L to cross the absolute cut-off. So, the imprecision of one TSH assay causes as many patients crossing the cut-off as using a different TSH assay.
4. Discussion

In the present study we performed a method comparison between three frequently used TSH immunoassays focusing on the low concentration range, as this range is important for clinical decision making in case of (subclinical) hyperthyroidism and thyroid cancer.

The studied TSH immunoassays showed standardization differences with a bias between 7% and 16%. When analyzing the standardization differences for the low range, the bias differed between 1% and 14%. Interestingly, the assays which were standardized against the same WHO standard varied more from each other than compared to the assay with a different WHO standard. The imprecision differed between 1.6% and 5.5% and increased in the lower range of the assay.

Several guidelines use absolute cut-off TSH values in the low range for clinical decision making. In the American and European guidelines for thyroid cancer and in the European guideline for subclinical hyperthyroidism an absolute TSH value of 0.10 mIU/L is mentioned as an important value above or below which recommendations are given concerning starting or adjusting medication and performing diagnostic testing. In order to establish whether the observed standardization differences translate into clinically relevant differences, we looked at all TSH results reported by our laboratory in the last six months. We calculated that 1.5% of all TSH measurements <0.4 mIU/L would theoretically lead to a different clinical outcome, concerning this absolute cut-off of TSH = 0.1 mIU/L. Naturally, not all the measurements around 0.1 mIU/L concerned patients with subclinical hyperthyroidism or thyroid cancer.

The imprecision using the Cobas assay had the same repercussions on the different classification of samples around TSH = 0.1 mIU/L as the greatest bias between different assays. For the Alinity and Atellica assays with a slightly higher imprecision for TSH <0.4 mIU/L the impact on the clinical management difference would be even higher than the immunoassay bias. Although the standardization differences for the total concentration range are evident, we consider these standardization differences too small to be clinically relevant in the low concentration range. Nonetheless, it is important to be cautious when interpreting concentrations close to a clinical cut-off value, especially if the methods are poorly standardized and the thresholds are not assay-specific.

The widespread use of arbitrary dichotomous cut-offs based on continuous data has limitations. Considering imprecision and bias, combined in total laboratory error or measurement uncertainty, are crucial when establishing clinical cut-offs. Total laboratory error or measurement uncertainty encompasses variability throughout the entire testing process, including pre-analytical, analytical, and post-analytical factors. Additionally, within-subject biological variation contributes to
variability in results and should also be considered when interpreting laboratory results. Both the total laboratory (analytical) variation and within-subject biological variation (estimated at 17.7% for TSH, source: https://biologicalvariation.eu/search?query=TSH accessed on 2 August 2023) contribute to the critical difference, also called the reference change value, the smallest difference between sequential laboratory results that is associated with a real change in the patient. By accounting for these factors, a better understanding of the limitations associated with arbitrary dichotomous cut-offs is achieved, enabling more informed and accurate interpretations of assay results.

Over time, a trend towards smaller standardization differences for TSH assays is seen in multiple studies. Comparing our results to other method comparisons is generally complicated by limited mention of the assay specifics. For some method comparisons, it is difficult to analyze results side by side because of different statistical methods, such as linear regressions instead of Passing-Bablok regressions or because of anonymization of the used assays. The imprecision is known to increase with lower concentration, which we also observed. The clinical management difference we observed is smaller than previous studies assessing the effect on clinical and subclinical hypothyroidism. This may be explained by the proportional difference on low vs high concentrations and improvements in bias between assays over time.

Taken together, we performed a method comparison of commonly used TSH assays and focused on the lower range of TSH concentrations. The bias in the lower concentration range showed a maximum of 14% between methods. Based on this proportional bias and the imprecision we estimated the theoretical effect on measured TSH concentrations in the last six months in our lab. This effect seems to have limited clinical consequences for the low TSH concentration range. Nevertheless, as guidelines use absolute TSH values to guide clinical decision making, caution must be applied with TSH values close to these cut-offs.

**Acknowledgements:** We would like to thank the technicians the laboratory for endocrinology, Amsterdam Medical Centers, Amsterdam, The Netherlands.

**Research funding:** None declared.

**Author contributions:** All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

**Competing interests:** Authors state no conflict of interest.

**Informed consent:** Not applicable

**Ethical approval:** Not applicable
**Data availability:** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.
5. References


10. Alexander EK, Pearce EN, Brent GA, Brown RS, Chen H, Dosiou C, Grobman WA, Laurberg P, Lazarus JH,


6. Legends of tables and figures

**Figure 1.** Passing-Bablok regression analysis for TSH, on the left for the total sample set (n=60) and on the right for TSH <0.4 mIU/L (n=36). A & B: Cobas assay vs Alinity assay. C & D: Cobas assay vs Atellica assay. E & F: Atellica assay vs Alinity assay. For all p<0.0001.

**Table 1.** Coefficient of variation (CV) for duplicate TSH measurements
<table>
<thead>
<tr>
<th>Assay</th>
<th>Range</th>
<th>CV (%)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cobas</td>
<td>&lt;0.4 mIU/L</td>
<td>3.0%</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>≥0.4 mIU/L</td>
<td>1.6%</td>
<td>8</td>
</tr>
<tr>
<td>Alinity</td>
<td>&lt;0.4 mIU/L</td>
<td>4.5%</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>≥0.4 mIU/L</td>
<td>3.5%</td>
<td>7</td>
</tr>
<tr>
<td>Atellica</td>
<td>&lt;0.4 mIU/L</td>
<td>5.5%</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>≥0.4 mIU/L</td>
<td>3.7%</td>
<td>7</td>
</tr>
</tbody>
</table>
A. Distribution samples for method comparison

B. Distribution samples for duplicate measurements