Abstract

Although pituitary thyrotropin (TSH) and thyroid hormones are physiologically interrelated, interpretation of measurements is conventionally done separately. Classification of subclinical thyroid dysfunction depends by definition solely on an abnormal TSH. This study examines a composite multivariate approach to disease classification.

Methods: Bivariate and trivariate reference limits were derived from a thyroid-healthy control group (n=271) and applied to a clinically diverse sample (n=820) from a prospective study, comparing their diagnostic efficiency with the conventional method.

Results: The following 95% reference limits were derived from the control group: (i) separate reference intervals for TSH, free thyroxine (FT4) and free triiodothyronine (FT3); (ii) bivariate composite reference limits for the logarithmically transformed TSH and FT4, and (iii) trivariate composite reference limits including all three parameters. A multivariate approach converts the “rectangular” or “cuboid” graphical representations of the independent parameters into an ellipse or ellipsoid. When applying these reference limits to the clinical sample, thyroid dysfunctions were classified differently, compared with the separate method, in 6.3 or 12% of all cases by the bivariate or trivariate method respectively. Of the established dysfunctions according to the separate intervals, 26% were reclassified to “euthyroid” by using the bivariate limit. Discrepancies from the laboratory-evaluated reference range were less pronounced.

Conclusions: Frequent divergencies between composite multivariate reference limits and a combination of separate univariate reference intervals suggest that statistical analytic techniques may heavily influence thyroid disease classification. This challenges the validity of the conjoined roles of TSH currently employed as both a sensitive screening test and a reliable classification tool for thyroid disease.

Introduction

Contemporary thyroid function testing relies heavily on the measurement of pituitary thyrotropin (TSH) (1, 2). The current definition of subclinical hypothyroidism or hyperthyroidism is solely biochemical, based on a TSH value outside the reference range, whereas thyroid hormones still remain within their respective reference limits. Using such a TSH-based definition, the prevalence of thyroid disease has been reported to be as high as 10% in the general population (3, 4, 5, 6). Despite this precise biochemically based definition, classification...
of thyroid dysfunction has been beset by considerable uncertainty, prompting a long-standing debate regarding the reference limits of TSH, which hitherto is unresolved (7, 8, 9, 10, 11).

Though the definition of the optimal reference limits for TSH has included many influences such as gender, age, ethnicity, and contamination by hidden disease, one important aspect has received scant attention (12, 13, 14, 15, 16, 17), namely the diagnostic potential of a composite reference interval of TSH with thyroid hormones. The reference ranges of TSH, free thyroxine (FT4), and free triiodothyronine (FT3) are potentially interdependent. Consequently, an acceptable all-encompassing reference interval cannot be confidently obtained merely by the summation of three univariate reference intervals. Hence, it should be better derived by a composite multivariate measure of all parameters and their interrelationships.

In previous communications, we have demonstrated that the controlled production of the three hormones TSH, FT4, and FT3 is an intricately entwined interactive process (18, 19). This is based on homeostatic principles that employ various feedback and feedforward motifs. Our studies have shown that individuals possess unique, closely controlled and limited set points for these parameters, which together collaborate to produce the optimum outcome for health (18, 20, 21). In thyroid insufficiency, this interlocking expression of control is significantly disturbed, leading to a new perspective of treatment modalities (19).

Accordingly, this has raised the question as to whether the various expressions of the parameters defining thyroidal health have implications for the efficiency of diagnostic procedures to distinguish normal function and dysfunction. We have therefore conducted more complex multifactorial examinations of the diagnostic efficiency of the test procedures, as opposed to the univariate and isolated use of single parameters, which is currently favored. This approach converts the “rectangular” graphical production of the interaction of the reference range of two parameters (e.g. TSH and FT4) into an elliptical form, whose diagnostic efficiency can now be coevaluated against the standard methods. This concept has already been considered (22, 23), but in the former early case (22), the study was conducted using the only available and more inaccurate approaches of the Free Thyroxine Index (FTI) and the first-generation TSH, and in the latter (23), only third-generation TSH and FT4 were considered.

Application of this novel approach also demands examination of more than one patient panel to justify its validity. In this study, we first established a composite reference system in a control group and, second, applied it to another sample from a recent prospective study (18). We then assessed the diagnostic efficiency of the composite referencing in this large and diverse clinical sample, and also compared the outcome with the routine laboratory-used reference method.

Methods

Patients

Samples were collected as part of a prospective study conducted by the Department of Nuclear Medicine at Klinikum Lüdenscheid, Lüdenscheid, Germany (www.ClinicalTrials.gov, NCT 01969552). The results of the primary study outcome on thyroid hormone equilibria have been previously reported (18). The original study involved 1912 consecutively seen adult patients, free of severe comorbidity. The study was approved by the local ethics committee, and all participants gave their written informed consent.

This study is a secondary analysis on reference limits involving two subgroups: (i) a thyroid-healthy and otherwise healthy control group of 271 subjects, who showed no clinical evidence for the presence of thyroid dysfunction based on patient history, symptoms, physical examination, antibody status, and thyroid imaging; and (ii) a consecutive series of 820 untreated patients from the prospective sample. The demographic characteristics for both groups are given in the “Results” section. Patients with major nonthyroidal illness, potentially interfering comorbidities such as pituitary disease, pregnancy, renal insufficiency, and on thyroid hormone replacement, were all ineligible or excluded. All patients included were seen as outpatients. Clinical information included details on patient history, thyroid medication, other medication, thyroid-associated symptoms, a thyroid-related physical examination, and a standardized questionnaire documenting gender, age, height, weight, smoking habits (75% answered), prior surgery, or radiiodine treatment. Laboratory tests routinely included FT3, FT4, and TSH, and, if autoimmune thyroiditis was suspected or to be excluded, thyroid peroxidase antibodies (TPO-Abs) or TSH receptor antibodies (TSH-R Abs). An ultrasound of the thyroid gland was routinely performed, whereas thyroid scintigraphy was restricted to patients with major
Clinical Study

R Hoermann and others

Composite reference range for thyroid hormones

174:6 | 737

ultrasound findings such as nodules with a diameter exceeding 1 cm. Lüdenscheid is located in a former area of mild iodine deficiency, where multinodular goiter, nonfunctioning thyroid nodules, and toxic adenomas are still prevalent in the local population.

Laboratory methods

TSH was measured with an automated direct chemiluminescence method (TSH3-Ultra, ADVIA Centaur XP, Siemens Healthcare). This third-generation assay for TSH was calibrated with the third International Standard of the World Health Organization for human TSH (IRP 81/565). It showed a range of linearity from 0.006 to 160.03 mIU/L and functional sensitivity of 0.008 mIU/L. Intra-assay coefficients of variation (CVs) in pooled serum samples in the range from 0.52 to 132.8 mIU/L (n=20) were 1.4 – 2.4%. Inter-assay imprecision measured in duplicate over 10 consecutive days was 0.9 – 2.9%. At a TSH value of 0.52 mIU/L, the intra-assay CV was 1.4%, the inter-assay CV 2.2%, but at the functional sensitivity (TSH of 0.008 mIU/L), the inter-assay CV rose to 14.1% (24).

FT3 and FT4 were measured with an automated competitive chemiluminescence method (ADVIA Centaur XP, Siemens Healthcare). The range of linearity was 0.3 – 30.8 (FT3) and 1.3 – 155 pmol/L (FT4). Serum samples with FT3 concentrations ranging from 2.9 to 14.2 pmol/L showed intra-assay CVs from 2.4 to 3.1% and inter-assay CVs from 2.3 to 3.9%. Serum samples with FT4 concentrations in the range of 9.3 – 38.8 pmol/L showed intra-assay CVs from 2.2 to 3.3% and inter-assay CVs from 2.5 to 4.0%.

Intra-individual CVs (biological variation) in duplicates (n=72, with TSH values ranging from 0.2 to 8.8 mIU/L) were 26% for lnTSH, 34% for the bivariate Mahalanobis distance, and 29% for the trivariate Mahalanobis distance.

TPO-Abs were measured with an automated competitive chemiluminescence method (Anti-TPO, ADVIA Centaur XP, Siemens Healthcare). TSH-R Abs were measured with a competitive ELISA (Anti-TSH Receptor, EUROMMUN AG, Lübeck, Germany). Thyroid function tests were performed by a single accredited institution, the Institute of Laboratory Medicine of Klinikum Lüdenscheid, Lüdenscheid, Germany. Standard laboratory quality procedures were routinely employed, and regular participation in interlaboratory tests was part of the quality management strategy.

The laboratory-evaluated reference intervals used for routine diagnostics were as follows: 0.4 – 4 mIU/L for TSH, 3.1 – 6.8 pmol/L for FT3, 10 – 23 pmol/L for FT4, <60 U/mL for TPO-Ab, and <2 mIU/L for TSH-R Ab.

**Figure 1**

Composite reference limit of (A) FT4 and natural lnTSH values and (B) FT3, FT4, and lnTSH in the control group (n=271). The circumference of the ellipse defines the bivariate 95% confidence limit of the composite interval, whereas the rectangular space shows the limits for the separate parameters. Outliers are depicted in the figures to show the complete sample but were not considered to define the reference intervals (see the Methods section). The respective trivariate reference limits are represented by the ellipsoid or cuboid in panel B.
Clinical Study
R Hoermann and others
Composite reference range for thyroid hormones

Table 1  Characteristics of the control group.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (s.d.) or Median (IQR) n=271</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (female/male)</td>
<td>203 (75%)/68 (25%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>46.6 (16.6), range 18–94</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>26.1 (4.57)</td>
</tr>
<tr>
<td>FT3 (pmol/L)</td>
<td>5.04 (0.55)</td>
</tr>
<tr>
<td>FT4 (pmol/L)</td>
<td>14.2 (1.73)</td>
</tr>
<tr>
<td>TSH (mIU/L)</td>
<td>1.50 (1.00; 2.05)</td>
</tr>
<tr>
<td>lnTSH (mIU/L)</td>
<td>0.38 (0.53)</td>
</tr>
<tr>
<td>Thyroid volume (mL)</td>
<td>12.1 (4.30)</td>
</tr>
</tbody>
</table>

Statistical methods

Descriptive data are shown as mean (standard deviation (s.d.)) or median (interquartile range (IQR)). Non-normally distributed TSH values were natural logarithmically transformed. In the control group, we obtained a reference interval for each single parameter, as defined by the 95% confidence interval after removing outliers detected by Horn’s method. We also determined a multivariate composite reference limit in the same control group, mathematically based on the Mahalanobis distance. As a prerequisite, multivariate normality between the logarithmically transformed TSH (lnTSH), FT4, and FT3 was confirmed by Henze-Zirkler’s multivariate normality test and quantile–quantile plot. The Mahalanobis distance represents a multidimensional generalization of the number of standard deviations by which a given point lies outside the multivariate center of the data, along the principal component axis. The measure is unitless and scale invariant, and takes into account the covariance of the data. Although the Mahalanobis distance itself relates to a unit circle, it can be readily rescaled and transformed into a two-dimensional ellipse or three-dimensional ellipsoid that defines the 95% confidence space of two or three parameters. The slight elliptical distortion (see Fig. 1A and B) from the ideal circle or globe reflects the amount of covariance that was present in the data. The limit for the Mahalanobis distance obtained in the control group was used as a cutoff to define multivariate normality in the independent second sample and create the graphical representation of the composite reference interval (ellipse or ellipsoid). We evaluated the diagnostic efficiency of the multivariate method against the standard procedure.

Table 2  Characteristics of the study group.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (s.d.) or Median (IQR) n=820</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (female/male)</td>
<td>623 (76%)/197 (24%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>49.4 (16.2), range 18–95</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>26.9 (5.27)</td>
</tr>
<tr>
<td>FT3 (pmol/L)</td>
<td>5.00 (4.70; 5.40)</td>
</tr>
<tr>
<td>FT4 (pmol/L)</td>
<td>14.1 (13.0; 15.5)</td>
</tr>
<tr>
<td>TSH (mIU/L)</td>
<td>1.38 (0.87; 2.15)</td>
</tr>
<tr>
<td>lnTSH (mIU/L)</td>
<td>0.32 (−0.14; 0.77)</td>
</tr>
<tr>
<td>Thyroid volume (mL)</td>
<td>14 (10; 19)</td>
</tr>
</tbody>
</table>

Diagnostic application of the composite reference limit to a consecutive series of untreated patients (n=820). The circumference of the ellipse (A) or the surface of the shaded ellipsoid (B) indicates the bivariate or trivariate reference limit derived from the control group respectively. Values outside the boundary of the ellipse (A) or ellipsoid (B) were regarded as “abnormal”. For comparison, reference intervals for the separate parameters are indicated by the rectangle (A) or cuboid (B).
Clinical Study: Composite reference range for thyroid hormones

R Hoermann and others

A two-sided P-value < 0.05 denotes statistical significance. We used the R statistical software base package (version 3.2.3 for Mac) and the added packages JGR 1.1-18, Deducer, 0.7-7, car 2.1-0, rgl 0.95.1201, robustbase 0.92.5, referenceIntervals 1.1.1, MVN 4.0, and chemometrics 1.3.9 (25, 26).

Results

Table 1 shows the characteristics of the control group. In this group of 271 subjects, we established conventional single reference limits for FT3, FT4, and lnTSH using the respective 95% confidence intervals for each parameter after outlier removal. The limits were 4.14–5.94 pmol/L for FT3, 11.06–17.31 pmol/L for FT4, and 0.47–3.77 mIU/L for TSH. In addition to the single-reference intervals, we derived the 95% composite bivariate reference interval for lnTSH and FT4, as defined by the Mahalanobis distance of 2.448, the cutoff for the 95% confidence area. Graphically, the multivariate approach converts the “rectangular” shape of the interaction of the reference range of two independent parameters into an ellipse that corresponds to the Mahalanobis distance at the 95% level taking into account their covariance (Fig. 1A). Similarly, we created the three-dimensional ellipsoid formed by the joined 95% space limit for FT3, FT4, and lnTSH (Fig. 1B). The critical Mahalanobis distance for all three parameters was 2.795.

The diagnostic efficiency of the multivariate method was evaluated against the standard procedure in a large independent series of untreated patients. The characteristics of this cohort comprising 820 patients are given in Table 2. Differences in the classification of abnormalities by use of univariate, bivariate or trivariate referencing for TSH, FT4 and FT3 originate from the values that, while lying outside the boundary of the ellipse or ellipsoid of the composite limits, fall within the boundaries of the rectangle or cuboid that defines the separate reference ranges or vice versa (Fig. 2A and B).

We compared the congruency or discrepancy in defining biochemical euthyroidism between the results of the composite or univariate referencing system (Table 3). We found an overall discrepancy rate of 6.3% between the conventional method and the bivariate approach, where both limits were taken from the healthy control group. The bivariate results were discrepant from the laboratory-used reference range in 5.9% of the cases (Table 3). For the trivariate reference range, the discrepancy rates were 12% from ranges derived from the control group and 8% from laboratory-established ranges. That means 5.9 or 8% of the classifications based on the laboratory-used reference range were unconfirmed by the composite bivariate or trivariate method respectively. Of the thyroid dysfunctions that were established by the univariate combination, 26% (42%) were reclassified to “euthyroid” by using the composite bivariate (trivariate) reference method (Table 3). The opposite situation in which abnormalities were reclassified from “euthyroid” to “dysfunctional” was much less frequently encountered (Table 3).

Discussion

Thyroid hormones and pituitary TSH are interrelated, and their relationships are determined by homeostatic
equilibria (19). Various feedback and feedforward motifs are jointly operative to produce the optimum levels in health and disease, and maintain T3 stability as long as it is physiologically possible (19, 27). This has raised the question as to whether this level of control impacts on the efficiency of diagnostic procedures, which accordingly might be improved to better distinguish normal function from thyroid dysfunction. The current widely used diagnostic approach favors the univariate use of TSH as a single statistical parameter, followed by a separate measurement of FT4 or FT3 when required (1, 2, 28). However, even where all three thyroid parameters are available, their interpretation is presently made in isolation. We have therefore conducted more complex multifactorial examinations of the diagnostic efficiency of test procedures. Physiologically, based on homeostatic principles and known hormonal interactions, a composite reference system seems both analytically and clinically more appropriate than employing three separate univariate reference ranges for each parameter. In this study, we have demonstrated considerable discrepancies in the disease classification by use of a composite reference limits, compared with the conventional approach.

Graphically, an ellipse represents the 95% bivariate reference interval for TSH and FT4, whereas a rectangular area describes the combination of the two single-reference intervals. The composite reference limit may be readily transformed into a single measure by projection on the unit circle. The corresponding quantile limit that comprises 95% of the data points can be numerically expressed by the radius of the corresponding circle, which is known as the Mahalanobis distance, or graphically displayed by the circumference of the ellipse, taking covariance into account. The observed slight elliptical distortion in the control group after removing statistical outliers from an ideal circle that could be expected in the absence of covariance may be readily explained by various influences such as age and BMI, as the lnTSH–FT4 relationship itself tends to be rather flat in the central euthyroid range (18, 23). The generation of the composite reference system in the control group in our study followed a similar approach as published by Kagedal et al. (22) and Ross et al. (23). Further attempts on interrelational definitions of “normality” have been reported by a few authors (29, 30). However, this approach has received scant attention in the literature and has never been adopted for routine thyroid function testing.

The disagreement on the reference range for TSH has fueled a fierce debate, and the subject has been addressed by numerous reviews (7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17). Improvements such as adjusting the reference limits for influences of ethnicity, age, body weight, and geographical region have been suggested (12, 13, 14, 15, 16, 17). However, aspects of multivariate interactive normality have not been considered in the recent discussion on reference limits. Nor have homeostatic principles been employed to biochemically define subclinical thyroid disease, which hinges by definition solely on an abnormal TSH (1, 2, 19, 28). The present data on the composite reference range reignite the debate by raising a fundamental question as to the reliability of present practice in diagnosis.

Our data demonstrate a considerable diagnostic disagreement between the univariate and multivariate reference limits, which were derived from the same control sample. In a diverse independent series of 820 consecutively seen patients, the methods gave rise to divergencies in the classification of euthyroid function versus thyroid dysfunction in 6.3% using the bivariate limit and in 5.9% using the trivariate limit. Remarkably, a substantial proportion of the dysfunctions that were classified as such according to the separate reference limits were reclassified to “euthyroid” by a composite limit, 26% by the bivariate limit, and 42% by the trivariate limit. The reclassification from “euthyroid” to “dysfunctional” was much less frequent.

It should be noted that the laboratory-evaluated reference intervals for the local population were generally wider than those obtained from the present control group (24). Their discrepancy rate to the composite limits was lower, but still substantial.

The present findings question the suitability of the isolated use of univariate reference intervals for disease classification. Given the high prevalence of subclinical thyroid disorders in the population, as defined solely by abnormal TSH measurements, the epidemiological and therapeutic implications are massive, especially as the divergences suggest that classification errors may arise purely from suboptimal statistical analytic techniques rather than clinical observation. Together with older data, our present findings should encourage broader exploration of the composite reference range toward its adoption for routine use.

It is time for the debate on the reference limits to show practical consequences and produce improvements in diagnostic efficiency. The classification of the thyroid disease should not rest on a single laboratory parameter, but rather consider the interactive involvement of all three parameters. As the thyroid gland operates far from its unstimulated resting point, a given FT4 level in the
euthyroid range contains inherent information about the impact of TSH stimulation, because TSH in turn reflects the thyroid hormone level. The two-dimensional or three-dimensional distribution of TSH, FT4, and, where applicable, FT3 clusters set points of healthy individuals. Therefore, it forms the physiological basis for the expression of multivariate normality that may be exploited for diagnostic purposes.

In addition to stating the respective values and reference limits of the single measurements, laboratories could report a distance from the multivariate population center or, alternatively, the percentile of the outlying position. This could provide a standardized quantitative measure of the parameters clinicians tend to jointly interpret. Provided the ellipse was constructed in a representative sample of the local population, the Mahalanobis distance should also facilitate harmonization of results from various assays that differ substantially in their absolute values.

Accordingly, the use of TSH as a sensitive screening test must be separated from its present role in unequivocally defining the biochemically euthyroid state. The latter should be derived from homeostatic principles that demand a composite multivariate measure of all parameters and their interrelationships, as opposed to the summation of separate univariate solutions. Although this relates to currently employed biochemical definitions of subclinical thyroid dysfunction, the clinical presentation of the patient must also be given a higher priority, including exploration of novel clinical markers of euthyroidism.

Although a composite reference system can be expected to make the classification and distinction between euthyroidism and thyroid dysfunction more reliable and better comparable among populations, TSH owing to its physiological role as an indirect and adaptive measure still displays a high index of individuality as defined by the ratio of intra-individual and inter-individual variation (24,31). This renders the adoption of population-based reference limits for individual decision making more difficult. Homeostatically derived individual set points independently of a population-based reference range have been proposed, but not clinically tested (19, 20, 21). We should note that the biochemical definition of subtle thyroid functional abnormalities does not invariably coincide with symptomatic disease and cannot decide by itself the necessity for treatment. Good clinical practice demands that any treatment decision is based on the synopsis of the patient history, symptoms, laboratory test and thyroid imaging.

Our study did not employ tandem mass spectrometry for measurement of free thyroid hormones, which is still impractical and too expensive for such large samples (32). However, known problematic situations such as major comorbidities and pregnancy were all excluded in our ambulatory samples.

A strength of the study is the use of well-characterized data sets and methods, one from a prospective study whose main outcome on homeostatic equilibria has been reported, as have further details pertinent to the performance and characteristics of the used assays in the local population (18, 24, 27, 33). In particular, the importance and relevance of additional FT3 measurement has been underscored by another study in the same local population demonstrating a u-shaped relationship of FT3, but none of FT4 or TSH, with a clinical outcome marker, mood disorders (34). Our results are in agreement with both theoretical expectations given known hormonal interactions and reported outcomes of a few studies that are available, particularly a large healthy collective reported by Ross et al. (23). Although we reused the data from a prospective study, the analysis was still secondary, and prospective clinical trials should be undertaken to prove the superiority of a composite reference range in a clinical setting.

In conclusion, this study reports on the derivation and diagnostic efficiency of a composite reference system for thyroid function testing. Frequent divergencies from the use of a combination of three separate univariate reference ranges for each parameter suggest that statistical analytic techniques may give rise to a substantial error rate in the widely accepted disease classification based solely on TSH abnormality. This is of particular concern in view of the high recorded prevalence of apparent subclinical thyroid disorders in the general population as defined by TSH (3, 4, 5, 6). Although its superiority awaits further proof by prospective trials, a composite approach appears to be more appropriately based on physiological principles and known hormonal interactions. Hence, the currently applied joint roles of TSH as both a sensitive screening test and a classification tool for thyroid disease should be separated, as the supporting diagnostic criteria differ in each situation.

Disclosure statement
J W D received funding and personal fees from Sanofi-Henning, Hexal AG, and Pfizer, and is co-owner of the intellectual property rights for the patent “System and Method for Deriving Parameters for Homeostatic Feedback Control of an Individual” (Singapore Institute for Clinical Sciences, Biomedical Sciences Institutes, Application Number 201208940-5, WIPO number WO/2014/088516). All other authors declare that there is no
Conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

Acknowledgment

The authors wish to thank Hans Günther Wahl, Institute of Laboratory Medicine, Klinikum Lüdenscheid, Lüdenscheid, Germany, for measurement of thyroid hormones.

References


14. Boucai L & Surks MI. Reference limits for serum TSH differ significantly between races and with age, potentially causing erroneous diagnosis in a substantial number of patients. *Clinical Thyroidology* 2009 21 10–12.


Received 12 January 2016
Revised version received 21 February 2016
Accepted 7 March 2016