Intra- and Interindividual Biological Variation of Five Analytes Used in Assessing Thyroid Function: Implications for Necessary Standards of Performance and the Interpretation of Results

Margaret C. K. Browning, R. P. Ford, S. J. Callaghan, and C. G. Fraser

Intra- and interindividual components of biological variation have been determined for total thyroxin (TT4), free thyroxin (FT4), total triiodothyronine (TT3), free triiodothyronine (FT3), and thyrotropin (TSH). Calculated analytical goals (CV, %) for the precision required for optimal patient care are: TT4 ≤ 2.5, FT4 ≤ 4.7, TT3 ≤ 5.2, FT3 ≤ 3.9, and TSH ≤ 8.1. The marked degree of individuality demonstrated for all hormones indicates that, if conventional population-based reference ranges are used uncritically, major changes in hormone concentration may not be correctly identified for some patients because observed values continue to lie within the reference range. At analyte concentrations approximating the mean values found in this study, and for analytical performance meeting the appropriate analytical goal, the differences required for consecutive results to be significantly different (p ≤ 0.5) have been calculated as: TT4, 14.7 nmol/L; FT4, 5.7 pmol/L; TT3, 0.6 nmol/L; FT3, 1.3 pmol/L, and TSH, 0.7 mili-int. unit/L.

Additional Keyphrases: population studies · reference interval · thyroxin · thyrotropin · triiodothyronine

Data on the biological variations of analyte concentrations or other quantities have important uses in clinical chemistry, including judging the usefulness of conventional population-based reference ranges (1), assessing the true significance of changes in results obtained for serial specimens from a single patient, and determining the standards of performance, or analytical goals, required to facilitate optimal patient care (2).

Thyroid dysfunctions are common endocrine disorders that can be readily and successfully treated. In mild forms, or at an early stage of disease, the clinical symptomatology may be somewhat nonspecific, commonly provoking lengthy biochemical investigation in many patients, only a few of whom will ultimately prove to have any abnormality of thyroid function. A diagnosis is made, or aided, by comparing the values observed for one or more relevant analytes with reference ranges that are usually derived from analyses of single specimens obtained from a large number of apparently healthy individuals. If treatment is instituted, the progress of individual patients is monitored by comparing their concentrations of serum hormone with those for the reference population, and with those obtained on previous occasions for the individual patient.

Five analytes—total thyroxin (TT4), free thyroxin (FT4), triiodothyronine (TT3), free triiodothyronine (FT3), and thyrotropin (TSH)—are generally measured, either singly or in various combinations, to diagnose and to monitor the efficacy of treatment of thyroid disease. Which analyte or analytes should be measured, particularly for the initial identification of patients with thyroid disease, is the subject of controversy (3-6); moreover, each analyte may be estimated by a plethora of methods, the performance characteristics of which vary widely. Few attempts have been made to establish objective analytical goals for precision for the analytes used to assess thyroid function. In two studies (7, 8), the values for analytical goals for TT4 that were considered necessary by clinicians for adequate patient care were widely divergent. However, standards of analytical performance so derived are based on subjective evidence, and formulation of analytical goals by collation and quantification of results of such surveys is likely to be unproductive and often misleading; moreover, the goals derived are likely to be strongly influenced by the quality of analyses currently available to the individual clinician (9). The state of the art as determined from the results of interlaboratory quality-assessment schemes has also been used to determine acceptable standards of assay performance (10-13). In this strategy (9), the arbitrary goals set are usually based on the performance achievable by the top 10% or 20% of laboratories participating; one cannot ascertain from such data, however, whether or not the standard of analytical performance is adequate, or even surpasses what is truly required for optimal care. Analytical goals based on studies of biological variation have been defined for several analytes but, of those used in the assessment of thyroid status, only TT4 (14) has been previously studied.

Some studies on the intra-individual variation of TT4 (15-18), TT3 (17, 18), FT4 (18), and TSH (19), and of interindividual variation of TT4 (15, 16) and TSH (19) concentrations in serum have been reported, but in no case have analytical goals for imprecision been defined from the small amounts of data available.

In this study, we have defined the contributions to the overall variance of analytical, intra-, and interindividual variance for each of the five analytes commonly used in the assessment of thyroid disease, and have defined the minimum analytical performance necessary for optimal patient care. Data on intra- and interindividual variation can also be used to define the usefulness of reference ranges in the identification of patients with disease. Previous studies (15, 16) of the biological variation of TT4 suggest that this hormone shows a high degree of individuality, which calls into question the appropriateness of using population-based

1 Nonstandard abbreviations: TT4, total thyroxin; FT4, free thyroxin; TT3, total triiodothyronine; FT3, free triiodothyronine; TSH, thyrotropin; VA, analytical variance; VB, intra-individual variance; VG, interindividual variance; VAD, analytical goal for variance; VQA, quality-assessment scheme goal for variance.
reference ranges to interpret results for individual patients. Here we have calculated the index of individuality for each of the five analytes used in assessing thyroid function.

Materials and Methods

Subjects. Twelve members of the laboratory staff (six men and six women, ages 22 to 46 years) were recruited for the study (Table 1). All the subjects were apparently healthy and none had a history of thyroid dysfunction. During the period of the study none took any medication or significant quantities of alcohol; smoking habits were nonremarkable. The subjects maintained their usual lifestyles during the study.

Specimen collection and handling. Ten 10-mL blood specimens were collected from each subject by conventional venepuncture, with minimal venous occlusion, between 08:30 and 09:30 hours at two- to five-day intervals within a five-week period. The same phlebotomist collected all the specimens, from seated subjects. The specimens were allowed to clot at room temperature, centrifuged at 3000 × g for 15 min at ambient temperature. We separated the serum, divided it into aliquots, and stored these at -40 °C until analysis. On the day of assay, one aliquot of each specimen from a single subject was thawed at ambient temperature and thoroughly mixed before analysis.

Analytical techniques. For each specimen we measured TT₄, FT₃, TT₃, FT₃, and TSH, as follows. TT₄ and TT₃ were measured by using in-house radioimmunoassays, with antisera obtained from the Scottish Antibody Production Unit (Law Hospital, Carluke, Scotland) and [¹²⁵I]thyroxine and [¹²⁵I]triiodothyronine purchased from New England Nuclear (Du Pont [U.K.] Ltd., Southamptom, England). In the TT₄ assay, antibody-bound hormone was separated by using polyethylene glycol, whereas in the TT₃ assay we used a double-antibody technique. FT₄ and FT₃ were measured with the FT₄ and FT₃ "Coat-A-Count" kits (Diagnostic Products [U.K.] Ltd., Wallingford, England). TSH was measured by time-resolved fluoroimmunoassay with the DELFIA reagent system and Arcus fluorometer (LKJ-Wallac, Turku 10, Finland).

The analytical protocol was designed to minimize analytical variance both within and between batches of analyses. All specimens from the same individual were analyzed in randomized duplicate within the same batch. Reagents, calibrators, and quality-control materials from the same sources were used for the analyses of all batches of each analyte. All analyses for a particular analyte were carried out by a single analyst and, except for TSH, were set up with a microprocessor-controlled sample distributor (Tecan 565; Laboratory Impex Ltd., Twickenham, Middlesex, England).

Calculation of results. Using analysis of variance techniques, we divided the total variance into the components attributable to analytical variance, intra-individual variance, and interindividual variance. Since all samples for any one individual were analyzed within the same batch, the analytical variance was calculated from the results of the duplicate analyses of each sample. Only the second value for each pair of duplicate results was used in the calculation of intra- and interindividual variances because using the mean of duplicate results would artificially reduce the analytical error by a factor of 0.71, i.e., by \( \sqrt{2} \) of replicates.

Results and Discussion

The mean and range of values obtained for TT₄, FT₃, FT₄, FT₃, and TSH for each of the 12 subjects are shown in Figure 1. We used these data to:

- derive analytical goals recommended for the analyses used in identifying and monitoring patients with thyroid disease,

- assess the usefulness of conventional population-based reference ranges in identifying patients with thyroid disease,

- assess the changes required in serial results before significance can be claimed.

Analytical Goals

It has been recommended (2, 14) that analytical variance should not exceed one-fourth of the relevant biological variance. The relevant biological variance will be either total biological variance or intra-individual variance, according to the clinical use to be made of the results. As a consequence of adoption of this recommendation, analytical variance will contribute less than 20% to the total observed variability. Using the data shown in Figure 1, we calculated for each of the five analytes, the overall mean value; the analytical (\( CV_A \)), intra-individual (\( CV_I \)), and interindividual (\( CV_D \)) variances; and the percentage of the total variance attributable to each component (Table 2). In each case, the analytical variance is \( <20\% \) of the total biological variance but \( >20\% \) of the intra-individual variance. The analytical goal to be achieved for identifying patients with disease and for serially testing an individual is (2):

\[
CV_A \leq \frac{1}{4} CV_I
\]

Table 1. Characteristics of the Subjects in the Study

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age, yr</th>
<th>Smoking habits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>Cigarette smoker (15/day)</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>Cigarette smoker (10/day)</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>Pipe smoker (70 g/week)</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>9</td>
<td>46</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>Nonsmoker</td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>Cigarette smoker (10/day)</td>
</tr>
</tbody>
</table>

The analytical performance achieved in this study is also shown. The only analysis that, performed on single samples, attains the analytical goal for precision is the TSH assay. The decrease in error that is achievable by performing replicate assays allows the analytical goal for FT₄ to be achieved when assays are performed in duplicate. The analytical performance of the methodology used to measure the remaining analytes is, by the criteria defined above,
inadequate for optimal patient care even when assays are performed in duplicate. Other approaches to defining acceptable standards of analytical performance depend upon subjective rather than objective criteria. Although the current "state of the art" may be assessed from the results of inter-laboratory quality-assessment schemes, from which one can attempt to define the performance that is currently achievable, one cannot assess objectively whether this "goal" is adequate for, or even surpasses, the standard required for clinical use. The organizers of the U.K. National External Quality Assessment Scheme for thyroid hormones accept much lesser precision as "tolerable" (13) than that defined from data on biological variation. For TT₄ and TT₃, a variability of bias of <5% is considered ideal but <10% is considered adequate. For FT₄, FT₃, and TSH, performance is considered to be inadequate only when the coefficient of variation of the bias exceeds 15%, 15%, and 25%, respectively. The difference between the analytical goals derived from

---

**Table 2. Mean of All Results and Calculated Average Components of Variance**

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Overall mean</th>
<th>Analytical</th>
<th>Intra-individual</th>
<th>Inter-individual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>V₀</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>TT₄</td>
<td>92 nmol/L</td>
<td>16.6</td>
<td>13.2</td>
<td>4.4</td>
</tr>
<tr>
<td>FT₄</td>
<td>19.2 pmol/L</td>
<td>1.3</td>
<td>13.0</td>
<td>5.9</td>
</tr>
<tr>
<td>TT₃</td>
<td>1.70 nmol/L</td>
<td>0.022</td>
<td>12.7</td>
<td>8.7</td>
</tr>
<tr>
<td>FT₃</td>
<td>5.4 pmol/L</td>
<td>0.17</td>
<td>9.3</td>
<td>7.6</td>
</tr>
<tr>
<td>TSH</td>
<td>1.34 milli-int. units/L</td>
<td>0.01</td>
<td>4.1</td>
<td>7.5</td>
</tr>
</tbody>
</table>

*Analytical (V₀), intra-individual (V₁), and inter-individual (V₀) variance, the percentage of each of total variance (V₀), and the coefficients of variation attributable to analytical (CV₀), intra-individual (CV₁), and inter-individual (CV₀) variance of results from 12 subjects.*

---

Fig. 1. Parametric means and absolute ranges for TT₄, FT₄, TT₃, FT₃, and TSH results from 12 subjects.
studies of biological variation and "acceptable standards of performance" as estimated from current analytical performance in external quality-assessment schemes serves to emphasize how much improvement in current analytical performance is needed.

Attempts have also been made to assess what clinicians regard as the standard of analytical performance required to generate clinically useful results. In two studies (7, 8), acceptable precision for \( \text{T}^4 \) was said to be 3.7% and 28.7%. This variation in what is considered acceptable probably reflects the quality of analytical service to which clinical staff had been exposed and the format of the clinical questions posed in such surveys. Moreover, clinicians probably include pre-analytical and biological variation in such estimates. Clearly, strategies for analytical goal setting that contain a subjective component, e.g., clinical opinion and assessment of the current state of the art, are less demanding than those derived from the objective evidence of data on biological variation. Although achieving these goals may be difficult with current methodology, no method is truly "good enough for clinical use" until it has attained them.

Usefulness of Conventional Population-Based Reference Ranges

Population-based reference ranges include both intra- and interindividual variation. The magnitude of the square root of the ratio of intra- to interindividual variance has implications for the most appropriate choice of reference values, because population-based reference ranges are of real value only when the intra-individual variation approximates the interindividual variation (20). If the intra-individual variation spans only a small part of the total range of values found for a population, major changes in analyte concentration for a particular individual may go undetected if such changes are interpreted only in the context of comparison with a conventional population-based reference range.

The square root of the ratio of intra-individual variance to interindividual variance and of the sum of analytical and intra-individual variances to interindividual variance have been calculated for each of the five analytes studied (Table 4). When analytical variance is included, all ratios are less than unity. Thus, in an individual patient, results outside the usual biological variation of that individual will not be recognized by the naive application of conventional population-based reference ranges (21). Removing the analytical variance emphasizes the insensitivity of population-based reference ranges to changes in individuals; the question must be posed as to how many patients with major alterations in thyroid function remain undetected when the results for only a single analyte are available and only a simple comparison with reference ranges is made. The use of multivariate tests (22)—e.g., the combination of \( \text{FT}_4 \) or \( \text{TT}_4 \) with TSH in suspected hypothyroidism, and of \( \text{FT}_4 \) or \( \text{TT}_4 \) with one of \( \text{FT}_3 \), \( \text{TT}_3 \), or TSH assays with low detection limit—may be more effective in diagnosing disease (22). Alternatively, the use of subject-specific reference values, which have been derived from previous investigations of the individual, may be a sensitive means of identifying changes in hormone concentrations. A major disadvantage of this approach in endocrine studies is that improvements in, and alterations to, hormone methodology may so profoundly affect the absolute values measured that longitudinal comparisons of data have limited use.

Significance of Serial Results from the Same Patient

The greater the analytical and intra-individual variation, the greater the difference must be between results from serial specimens before they can be said to be significantly different.

For \( p < 0.05 \), the difference required for significance is \( \pm 2.5 \sqrt{V_1} \). Table 5 shows the difference required between successive results for this level of significance to be attained as calculated, first, from the data on \( V_A \) and \( V_1 \) from this study; secondly, from the analytical variance equal to the analytical goal \( (V_{A0}) \), calculated from intra-individual variance; and thirdly, from the criteria for minimum acceptable performance in the U.K. National External Quality Assessment Scheme for thyroid hormones, with use of the mean value for each analyte from this study to convert the coefficient of variation into variance \( (V_{QA}) \). The data used in these calculations are for values found in healthy individuals. For values below and above the reference limits, the precision of the analytical method may be less, particularly for ligand binding assays. As yet no data are available on the biological variation of thyroid hormones in patients with thyroid disease. The data given in Table 5 emphasize the importance of using assays with high precision if patients are to be followed sequentially. However, for analytes with a high intra-individual variation, one must appreciate that, even if the precision is good, relatively large differences between the results from sequential specimens will be required before they can be said to be significantly different.

In conclusion, the data derived in this study demonstrate the following:

- analytical goals derived from data on biological variation for \( \text{TT}_4 \), \( \text{FT}_4 \), \( \text{TT}_3 \), \( \text{FT}_3 \), and TSH indicate that the standard of analytical performance generally achieved at present may not be adequate for ideal patient care;
- population-based reference ranges are of limited value for the interpretation of results of measurements of \( \text{TT}_4 \), \( \text{FT}_4 \), \( \text{TT}_3 \), \( \text{FT}_3 \), and TSH in that major changes in an individual may go undetected because of the high degree of intra-individual variance for these hormones; and
* the large intra-individual variation exhibited by TT4, FT4, TT3, FT3, and TSH means that, even when analytical goals for precision are met, relatively large differences between sequential results are required before two values can be said to be significantly different; the use of methodology with poor precision will greatly magnify the differences required to reach significance.

We thank all colleagues who participated in this study, Dr. Allan Prentice for taking the blood specimens, and Mrs. Margaret Cooper for typing the manuscript. The generous support of LKB-Wallac, Diagnostic Products (UK) Ltd., and Tayside Health Board is gratefully acknowledged.

References
