Toward a consensus on reference values for thyroid volume in iodine-replete schoolchildren: results of a workshop on inter-observer and inter-equipment variation in sonographic measurement of thyroid volume

M B Zimmermann¹, L Molinari², M Spehl³, J Weidinger-Toth⁴, J Podoba⁵, S Hess¹ and F Delange⁶

¹Swiss Federal Institute of Technology, Zurich, Switzerland, ²Children's Hospital, Zurich, Switzerland, ³University Hospital Saint-Pierre, Brussels, Belgium, ⁴Burgerspital, Solothurn, Switzerland, ⁵Post-Graduate Medical School, Bratislava, Slovakia and ⁶International Council for the Control of Iodine Deficiency Disorders, Brussels, Belgium

(Correspondence should be addressed to M B Zimmermann, The Laboratory for Human Nutrition, Swiss Federal Institute of Technology Zurich, Seestrasse 72/Postfach 474, CH – 8803 Rüschlikon, Switzerland; Email: michael.zimmermann@ilw.agrl.ethz.ch)

Abstract

Objective: Interpretation of thyroid ultrasonography for assessing goiter prevalence requires valid reference criteria from iodine-sufficient populations. Reports have suggested the current reference criteria for thyroid volume (Tvol) of WHO/ICCIDD (International Council for the Control of Iodine Deficiency Disorders) may be too high. Our objective was to determine if inter-observer and/or inter-equipment variability contributes to the disagreement in sonographic Tvol in children reported from iodine-sufficient areas.

Design: A 2-day workshop in which four experienced ultrasound examiners from around Europe measured Tvol in 45 6–12-year-old Swiss schoolchildren using four different portable ultrasound machines. One of the participating examiners (observer A) had generated the Tvol data in European children that are the basis for the WHO/ICCIDD reference criteria.

Methods: Sonographic Tvol was measured in each child by all four examiners on all four machines. Six hundred and eighty-four examinations were completed, with examiners having no knowledge of one another's results. Inter-observer and inter-equipment variation was calculated.

Results: Mean inter-equipment variation in Tvol was 15.2% (95% CI: 14.1, 16.3%). There were no significant differences in Tvol between equipment (P = 0.51). For all observers, the mean inter-observer variation in Tvol was 25.6% (95% CI: 23.9, 27.2%). At all ages and all body surface areas, there was a large systematic measurement bias (+30% volume) between the mean Tvol of observer A and the mean Tvol of observers B, C and D. Reanalysis using data from observers B, C and D reduced the mean inter-observer variation in Tvol to 13.3% (95% CI: 11.9, 14.7%). A correction factor for the systematic difference of operator A for the P50 and P97 of Tvol was estimated using analysis of covariance. When applied to the WHO/ICCIDD reference data, it sharply reduced the discrepancy between the WHO/ICCIDD criteria and those from other iodine-sufficient children around the world.

Conclusions: Inter-equipment error contributes minimally to reported differences in sonographic Tvol. Even among experienced examiners, inter-observer variation in sonographic Tvol in children can be high, and probably contributes to the current disagreement on normative values in iodine-sufficient children. A systematic bias at least partially explains why the WHO/ICCIDD reference data differ from those reported from other iodine-sufficient children around the world. The findings argue strongly for the standardization of methods used for sonographic measurement of Tvol in children.

European Journal of Endocrinology 144 213-220

Introduction

Goiter prevalence in school-age children is an important indicator for assessing iodine deficiency disorders (IDD) in a population. Sonographic measurement of thyroid volume (Tvol) is more precise than inspection and palpation for determination of goiter prevalence (1). Interpretation of thyroid ultrasonography requires valid reference criteria from iodine-sufficient populations, and the World Health Organization and International Council for the Control of Iodine Deficiency Disorders (WHO/ICCIDD) have proposed normative values for Tvol in children aged 6–15 years (2). However, several reports have suggested these reference criteria may be too high. Tvol values in iodine-sufficient US children (3), Swiss children (4) and Malaysian
Subjects and methods

Subjects

We randomly selected 45 healthy children (23 boys, 22 girls) from school records in Jona, a town 45 km from Zurich. The sample included six to eight children at each 1 year increment from 6 to 12 years; mean age was 9.2 years. This age range was chosen because it is recommended by WHO/UNICEF/ICCIDD for monitoring Tvol in children in the context of IDD control programs (2). To ensure the children had had a lifetime of iodine sufficiency, the sample included only children born and raised in Switzerland. Salt iodization began in Switzerland in 1922, was national in 1952 and the level of iodization has been 15–20 mg iodine/kg since 1980. Consequently, Swiss children today are likely to have had a steady and sufficient iodine intake since birth (10). Children from this community were sampled in a 1999 national iodine survey (4). That study found a median urinary iodine concentration of 115 μg/l (123 μg/g creatinine) among children aged 6–12 years, and no cases of goiter. Informed written consent was obtained from the school board and the parents of the children.

Methods

The children were registered and height and weight were measured (11). For the measurements, subjects removed their shoes, emptied their pockets and wore light indoor clothing. We chose four portable ultrasound machines in current use: (1) an Aloka SSD-500 with a 7.5 MHz 4 cm linear transducer; (2) an Aloka SSD-900 with a 7.5 MHz 4 cm linear transducer (Aloka, Mure, Japan); (3) a Toshiba JustVision-200 with a 8 MHz 5 cm linear transducer (Toshiba, Tokyo, Japan); and (4) a Siemens Sonoline with a 7.5 MHz 6 cm linear transducer (Siemens, Munich, Germany) located in a Thyromobil Van (Merck, Darmstadt, Germany). Four expert ultrasound examiners from around Europe were invited to participate. Two were radiologists (observers B and C), and two were endocrinologists with experience of using thyroid ultrasound in goiter surveys (observers A and D). Observer A generated the Tvol data that form the basis for the current WHO/ICCIDD normative thyroid volumes (2).

Sonographic Tvol was measured in each child by all four examiners on all four machines. Each examiner was assigned to an ultrasound machine and the children proceeded through the four stations. The examiners then moved to a new machine, the order of the children was shuffled, and the children proceeded again through the four stations. On the two Aloka machines and the Toshiba, subjects were measured sitting upright. In the Thyromobil van using the Siemens, they were measured in the supine position with a towel roll under the shoulders. There was an interval of at least five subjects between repeat examinations on a child by the same observer. It was planned that each child would undergo 16 ultrasound examinations, for a total of 720 examinations. Because of time constraints, 684 examinations were completed. At each ultrasound machine, an assistant recorded all measurements obtained by the observers, and the ultrasound stations were widely spaced. Therefore, examiners had no knowledge of one another’s results. There was no discussion of ultrasound methods and technique between examiners before or during the workshop. All measurements were done over 2 days.

Body surface area (BSA) was calculated using the formula: BSA = weight (kg)0.425 × height (cm)0.725 × 71.84 × 10−4 (2). Tvol was obtained using the method of Brunn et al. (12). For each lobe, the maximum perpendicular anteroposterior (AP) dimension and mediolateral (ML) dimension were measured on a transverse image of the largest diameter. The maximum cranio-caudal (CC) diameter was measured on a longitudinal image. The volume of each lobe was calculated as: AP diameter × ML diameter × CC diameter × 0.479 and the lobe volumes were summed (2, 12). The inter-observer and inter-equipment variations were calculated as follows: for each child on each machine, the mean of the four volumes (or diameters) produced by

www.eje.org
the four observers was defined as the ‘true’ volume or diameter. Then, for each child on each machine, the volume or diameter measured by one observer was subtracted from the volumes or diameters of the other observers (A–B, A–C, A–D, B–C, B–D, C–D). Finally, the average of the absolute differences was divided by the ‘true’ thyroid volume or diameter and expressed as a percentage, providing an individual inter-observer error for a specific machine. The inter-observer variation was expressed as the mean (S.E.) inter-observer error, averaged over all subjects and machines. To obtain the inter-equipment variation, a similar calculation was done using the volumes and diameters obtained from each child measured by each observer on the same machine.

**Statistics**

Data processing and statistics were done using SPLUS 2000 (Mathsoft, Seattle, WA, USA) and Excel (Microsoft, Seattle, WA, USA). Repeated measures ANOVA was used for testing differences among operators and equipment and modeling age, BSA and sex dependence. Spearman rank correlations were calculated. Inter-observer and inter-equipment variation were expressed as means, and 95% confidence intervals (CI) were calculated as mean ±1.96 × S.E. Analysis of covariance by operator, with age and BSA as covariates, was used to estimate a correction factor for the systematic difference in total Tvol between operator A and the other three operators.

**Results**

As shown in Table 1, the mean inter-observer variation in the measurement of the ML, AP and CC diameters ranged from 15.6 to 17.5%. For all observers, the mean inter-observer variation in total Tvol was 25.6% (95% CI: 23.9, 27.2%). There was no correlation between total Tvol and inter-observer variation (P = 0.9). There was a large and highly significant difference between the mean Tvol of observer A and the mean Tvol of observers B, C and D (P < 0.00001), but only minor differences between observers B, C and D (Fig. 1). The mean Tvol of observer A was 30% higher than those of the other observers, essentially due to systematic measurement of greater CC diameters by observer A (Table 1). The systematic bias in the measurements of observer A compared with the other three observers was present at all ages (Fig. 2) and all BSAs (Fig. 3). Therefore, we re-analyzed the data using only the measurements of observers B, C and D. As shown in Table 1, this reduced the mean inter-observer variation in Tvol to 13.3% (95% CI: 11.9, 14.7%). This reduction in inter-observer variation in Tvol was due to reduced variation in the measurement of the CC diameter.

As shown in Table 2, the mean inter-equipment variation (which includes intra-observer variation) in the three diameters was greatest for the AP dimension. Mean inter-equipment variation in total Tvol was 15.2% (95% CI: 14.1, 16.3%). There were no significant differences in Tvol between equipment (P = 0.51). Overall, mean Tvol (S.D.) for the Aloka SSD-500, Aloka SSD-900, Siemens Sonoline and Toshiba JustVision were 3.22 (1.22), 3.07 (1.12), 3.26 (1.27) and 3.19 (1.34) ml respectively. The mean CC diameters measured with the Siemens Sonoline were greater by 0.1–0.15 cm compared with those of the other three machines (P < 0.0005) (data not shown). There was a significant, but small, observer–equipment interaction (P = 0.0003), indicating the inter-equipment differences were not the same for all observers on all machines.
Discussion

There are several potential sources of inter-observer and/or inter-equipment error in thyroid ultrasound. First and foremost, the method for sonographic determination of T\textsubscript{vol} has not been clearly standardized. The ‘rotational ellipsoid’ method of Brunn et al. (12) with an optimized correction factor of 0.479 is the most widely used, but the original paper has only a cursory description of the method. Other investigators have used the ellipsoid method but with a correction factor of 0.52 (\(\pi/6\)) (8). Brunn et al. (12) state that the potential contribution of the thyroid isthmus to total gland volume should be ignored. However, others recommend that if the isthmus is >1 cm (whether this is width and/or depth is not specified), its volume should be included (13, 14).

Secondly, thyroid ultrasound is subjective in that finding and measuring the maximum diameters requires judgment and experience. Thirdly, there is no agreement on whether the subject should be measured in the supine or sitting position. The supine position may produce greater neck extension that could increase the CC diameter. Finally, a number of portable ultrasound machines are available and come equipped with a variety of transducers. Although higher frequency transducers (7.5–10 MHz) provide better resolution and are therefore recommended for thyroid scanning (15), lower resolution 5 MHz transducers are still commonly used (3, 16).

Despite this, previous reports have described generally small inter-observer variation in the sonographic measurement of T\textsubscript{vol} in children. In Malaysian children aged 7–10 years, Foo et al. found an inter-observer
error (S.D.) in ultrasound measurement of $T_{\text{vol}}$ of 3.4 (3.7)% (5). In Italian children aged 6–14 years, Vitti et al. reported an inter-observer error of 4.2% (mean) and 5.2% (median) (8). Hess & Zimmermann (4), in Swiss children aged 6–12 years, found an inter-observer error (S.D.) of 3.7 (3.5)%; Özgen et al. reported an inter-observer error (S.D.) of 13.3 (8.2)% in 7–16-year-old Turkish children (9). The small inter-observer variation in these studies is not surprising, in that they compare two or three examiners who typically share methods, training and equipment.

The present study brought together experienced ultrasound examiners from Switzerland, Belgium and Slovakia. They had different backgrounds and training (two were radiologists, two were endocrinologists) and they had not previously met. The observers purposely avoided discussion of ultrasound methods or technique before or during the workshop, and during the workshop had no knowledge of one another’s results. Therefore, our results are likely to be unbiased and probably reflect true variation in sonographic measurement of $T_{\text{vol}}$.

For all observers, the 95% CI for inter-observer variation in $T_{\text{vol}}$ was 23.9–27.2% (Table 1). This suggests differences of $<27.2\%$ in sonographic measurement of $T_{\text{vol}}$ by separate observers in school-age children should be considered statistically not significant. However, as shown in Figs 2 and 3, this high variability was caused by a systematic bias in the measurements of observer A, apparently caused by small differences in technique and estimation of thyroid anatomy. Observer A appeared to apply the transducer to the neck with greater pressure than the other observers. The resulting compression and flattening of the thyroid may have increased the ML diameter and slightly increased the CC diameter (Table 1). Whereas observers B, C and D nearly always measured the CC diameter on a single scan, observer A more often determined the CC diameter by measuring a portion of the diameter on one scan, the remainder of the diameter on a second scan and then adding the lengths. This approach may have contributed to the larger CC diameters of observer A compared with the other three observers, and increased the overall variation in $T_{\text{vol}}$ (Table 1).

In discussions after the workshop, it became clear that observers B, C and D shared very similar methods and technique for thyroid ultrasound. Because of the systematic bias in the measurements of observer A, we re-analyzed the data using only the measurements of

<table>
<thead>
<tr>
<th>Inter-equipment/Intra-observer variation (as percentages)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left lobe</strong></td>
</tr>
<tr>
<td>Volume</td>
</tr>
<tr>
<td>Mediolateral</td>
</tr>
<tr>
<td>Anteroposterior</td>
</tr>
<tr>
<td>Craniocaudal</td>
</tr>
</tbody>
</table>

| **Right lobe**              |
| Volume                      | 19.8 (18.3, 21.2) |
| Mediolateral                | 11.8 (10.9, 12.7) |
| Anteroposterior             | 15.1 (13.8, 16.5) |
| Craniocaudal                | 10.2 (9.4, 11.0)  |

* Values are means (95% CI).
observers B, C and D (Table 1). This sharply reduced the inter-observer variation in Tvol to 13.3% (95% CI: 11.9, 14.7%). In contrast to the result with all observers, this suggests there is a 95% probability that a measurement of total Tvol by one observer will be within ±14.7% of the measurement performed by a second similarly trained observer.

The mean inter-equipment/intra-observer variation in total Tvol was 15.2% (95% CI: 14.1, 16.3%) (Table 2). This suggests there is a 95% probability that a measurement of Tvol by an experienced observer will be within ±16.3% of a repeat measurement by the same observer using different equipment. The inter-equipment variation may have been at least partially due to differences in transducer frequency (7.5 vs 8 MHz) and transducer length (4 vs 5 vs 6 cm). Because the children were examined supine with the Siemens and sitting with the other three machines, it is possible the greater mean CC diameters produced by the Siemens could have been at least partly due to greater extension of the neck in the supine versus the sitting position. However, this difference, although statistically significant, is unlikely to be of clinical relevance, as there were no significant differences in total Tvol between equipment (P = 0.51) and the largest difference in mean total Tvol, between the Aloka SSD-900 and the Siemens Sonoline, was only 0.15 ml at a Tvol of just over 3 ml.

As shown in Figs 2 and 3, there was a clear inter-subject difference at all ages and BSAs between the Tvol of observer A and the remaining observers. Because observer A generated the Tvol data in European children that are the basis for the WHO/ICCIDD normative criteria, this systematic bias at least partially explains why the WHO/ICCIDD reference data are greater than those reported from other iodine-sufficient children around the world (3–5, 13). It also explains the discrepancy in sonographic Tvol recently reported by two different investigative teams in Swiss children (4, 7), one of which included observer A (7).

To eliminate this systematic bias from the WHO/ICCIDD normative Tvol data, we estimated a correction factor for the systematic difference of operator A for the P50 and P97 of Tvol using analysis of covariance by operator, with age and BSA as covariates. We found small differences between boys and girls, consistent with known sex differences; these, however, did not achieve statistical significance. The correction factor was sex independent and on the logarithmic scale, age and BSA independent, and was the same for age (6–12 years) and BSA (0.8–1.5 m²).

A provisional value for the correction factor on log scale is -0.346 (s.e. = 0.025); in other words, the usual calculation of the P97 centile, P97(age) = exp(mean(age) + 1.96 × s.d.(age)), with mean and s.d. on log scale and exp(\(x\)) = \(e^x\) (the exponential of \(x\)), should be changed to P97(age) = exp(mean(age) - 0.346 + 1.96 × s.d.(age)), and similarly with respect to BSA. This is equivalent to multiplying the current WHO/ICCIDD values by exp(-0.346) = 0.71.

Table 3 compares the WHO/ICCIDD normative data, the same data adjusted with the correction factor generated in the present study, and four studies of sonographic thyroid gland volume in iodine-sufficient boys as a function of age.

<table>
<thead>
<tr>
<th>Source</th>
<th>Age (years): 6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
</tr>
<tr>
<td>WHO and ICCIDD (2)</td>
<td>3.2</td>
<td>5.4</td>
<td>3.4</td>
<td>5.7</td>
<td>3.7</td>
<td>6.1</td>
<td>4.1</td>
</tr>
<tr>
<td>WHO and ICCIDD Corrected</td>
<td>2.3</td>
<td>3.8</td>
<td>2.4</td>
<td>4.0</td>
<td>2.6</td>
<td>4.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Gutekunst &amp; Martin-Teichert (13) (Europe)</td>
<td>1.5</td>
<td>3.5</td>
<td>1.8</td>
<td>4.0</td>
<td>2.0</td>
<td>4.5</td>
<td>2.4</td>
</tr>
<tr>
<td>Hess &amp; Zimmermann (4) (Switzerland)</td>
<td>2.1</td>
<td>4.1</td>
<td>2.4</td>
<td>4.5</td>
<td>2.8</td>
<td>5.0</td>
<td>3.1</td>
</tr>
<tr>
<td>Xu et al. (3) (USA)</td>
<td>2.2</td>
<td>3.2</td>
<td>2.6</td>
<td>4.0</td>
<td>2.8</td>
<td>4.7</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Values are the 50%tile (P50) and upper limit of normal (97%tile (P97)) for sonographic thyroid volume.
Table 4 Comparison of the WHO/ICCIDD normative data, the same data adjusted with the correction factor generated in the present study, and four studies of sonographic thyroid gland volume in iodine-sufficient boys as a function of body surface area (BSA).

<table>
<thead>
<tr>
<th>Source</th>
<th>BSA (m²)</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>1.4</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
<td>P97</td>
<td>P50</td>
</tr>
<tr>
<td>WHO and ICCIDD (2)</td>
<td>3.0</td>
<td>4.7</td>
<td>3.4</td>
<td>5.3</td>
<td>3.8</td>
<td>6.0</td>
<td>4.4</td>
<td>7.0</td>
<td>5.0</td>
</tr>
<tr>
<td>WHO and ICCIDD Corrected</td>
<td>2.1</td>
<td>3.3</td>
<td>2.4</td>
<td>3.8</td>
<td>2.7</td>
<td>4.2</td>
<td>3.1</td>
<td>5.0</td>
<td>3.5</td>
</tr>
<tr>
<td>Foo et al. (5) (Malaysia)</td>
<td>2.4</td>
<td>4.2</td>
<td>2.7</td>
<td>4.6</td>
<td>2.9</td>
<td>5.1</td>
<td>3.1</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>Hess &amp; Zimmermann (4) (Switzerland)</td>
<td>2.1</td>
<td>3.3</td>
<td>2.3</td>
<td>3.6</td>
<td>2.6</td>
<td>4.1</td>
<td>3.0</td>
<td>4.7</td>
<td>3.5</td>
</tr>
<tr>
<td>Xu et al. (3) (USA)</td>
<td>2.5</td>
<td>3.8</td>
<td>2.8</td>
<td>4.3</td>
<td>3.1</td>
<td>4.8</td>
<td>3.4</td>
<td>5.3</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Values are the 50%tile (P50) and upper limit of normal (97%tile (P97)) for sonographic thyroid volume.

data, it sharply reduces the discrepancy between the WHO/ICCIDD criteria and those generated in other groups of iodine-sufficient children from around the world.

Our findings illustrate several important points. First, inter-equipment error appears to contribute only minimally to differences in $T_{vol}$ measurement among groups working in the field of IDD around the world. Second, inter-observer variation in the sonographic measurement of $T_{vol}$ in children can be high, even among experienced examiners. Small differences in sonographic technique may introduce significant systematic bias into measurements of $T_{vol}$. Third, the current disagreement on normative values for $T_{vol}$ by ultrasound in iodine-sufficient children may be due in part to inter-observer variation. Finally, the narrow inter-observer variation among the three examiners who shared similar technique is encouraging in that it suggests standardization of ultrasound technique could increase the precision of $T_{vol}$ measurement among groups working in the field of IDD around the world. These data argue strongly for the standardization of methods used for ultrasonography in the measurement of thyroid volume in children.

Acknowledgements

We thank WHO, and especially Dr B de Benoist, Micronutrient Focal Point, Dept. of Nutrition for Health and Development of WHO, ICCIDD, and the Human Nutrition Laboratory of the Swiss Federal Institute of Technology for organization and financial support of the workshop. Also, we thank Merck KaA Darmstadt, and especially Dr U Hostalek, for providing the Thyromobil van, and Aloka and Toshiba for the ultrasound equipment and technical support. Finally, we thank B Bussmann, the Jona School Board, and the children for their participation.

References


Received 23 August 2000
Accepted 13 November 2000