Prospective evaluation of thyroid imaging reporting and data system on 4550 nodules with and without elastography

Gilles Russ, Bénédicte Royer, Claude Bigorgne, Agnès Rouxel, Marie Bienvenu-Perrard, and Laurence Leenhardt

1Centre of Pathology and Radiology, 14 Avenue René Coty, 75014 Paris, France, 2Department of Nuclear Medicine, Pitie-Salpétrière Hospital, Pierre et Marie Curie University, Boulevard de l’Hôpital, 75651 Paris, France, Departments of 3Pathology and 4Nuclear Medicine, Cochin Hospital, 27 Rue du Faubourg Saint Jacques, 75014 Paris, France

(Correspondence should be addressed to G Russ at Centre of Pathology and Radiology; Email: gilles.russ@wanadoo.fr)

Abstract

Objective: To evaluate prospectively the diagnostic accuracy of the thyroid imaging reporting and data system (TI-RADS) and its interobserver agreement and to estimate the reduction of indications of fine-needle aspiration biopsies (FNABs).

Design: A prospective comparative study was designed.

Methods: In two years, 4550 nodules in 3543 patients were prospectively scored using a flowchart and a six-point scale and then submitted to US-FNAB. Results were read according to the Bethesda system. Histopathological results were available for 263 cases after surgery. Sensitivity, specificity, negative predictive value (NPV) and positive predictive value, and accuracy were calculated for the gray-scale score, elastography, and a combination of both methods. Interobserver agreement was calculated using the kappa statistic. The reduction in the number of FNABs was estimated.

Results: When compared with cytopathological results, sensitivity, specificity, NPV, and accuracy were 95.7, 61, 99.7, and 62% for the TI-RADS gray-scale score; 74.2, 91.1, 98, and 90% for elastography; and 98.5, 44.7, 99.8, and 48.3% for a combination of both methods respectively. When compared with histopathological results, the sensitivity of the gray-scale score, elastography, and a combination of both methods were 93.2, 41.9, and 96.7% respectively. Interobserver agreement for the six-point scale and the recommendation for biopsy were substantial (κ value = 0.72 and 0.76 respectively). The reduction in the number of FNABs was estimated to be 33.8%.

Conclusion: The TI-RADS score has high sensitivity and NPV for the diagnosis of thyroid carcinoma. A hard nodule should always be considered as suspicious for malignancy but elastography cannot be used alone. Combination of elastography with gray-scale can be used to improve sensitivity or specificity. Interobserver agreement and decrease in unnecessary biopsies are significant.

Introduction

During 2009, two separate teams (1, 2) suggested an interesting thyroid imaging reporting and data system (TI-RADS) derived from the breast imaging reporting and data system (BI-RADS) (3). Initial TI-RADS scores stratified the risk of thyroid nodule malignancy with ultrasound (US) scanning. However, they appeared difficult to apply in daily practice.

With this limitation in mind, our group built a reproducible TI-RADS system that could be used by many institutions. First, we proceeded to perform a retrospective study on 500 nodules (4). The sensitivity, specificity, and odds ratio of each US sign were calculated, and a specific vocabulary and a standardized report were established. A flowchart was developed to easily define the score of a particular nodule. Sensitivity, specificity, and odds ratio of this version of the TI-RADS score were 95, 68, and 40 respectively (95% CI, 33.7–47.9) (4).

The main goal of this study was to test prospectively the diagnostic accuracy of our TI-RADS score with this flowchart by comparing gray-scale score, elastography, and a combination of both methods with cyto- and histopathological results. The three imaging methods were also intercompared. Secondary objectives were to determine interobserver agreement and to estimate the decrease in unnecessary fine-needle aspiration biopsies (FNABs).

Materials and methods

Imaging technique and TI-RADS scoring

Between April 2010 and March 2012, 3543 patients were referred to our specialized thyroid clinic for US-guided FNABs (US-FNAB) of 4550 nodules. US scanning and sono-elastography were performed using...
Thyroid FNAB and cytological interpretation techniques

All 4550 nodules were submitted to FNAB and biopsies were performed using a capillary US-guided FNAB technique with 27 G needles. In most of the cases, only one needle pass was made per lesion. The Bethesda classification was used to interpret smears (8). Cytological specimens were divided into six categories ranging from 1 to 6, respectively, as follows: nondiagnostic, 91 cases (2%); benign, 3518 cases (77.3%); atypia of undetermined significance, 415 cases (9%); suspicious for a follicular neoplasm, 256 cases (5.6%); suspicious for malignancy, 130 cases (2.8%); and malignant, 140 cases (3.1%).

Histopathological results

Histopathological results were available for 263 cases after surgery for the gray-scale score and for 74 inclusive cases where elastography was performed as well. Surgery was indicated based on cytopathological results (69 malignant, 86 suspect, 89 follicular neoplasm, and nine atypia of indeterminate significance cases) or when the nodule was benign but larger than 3 cm and causing compressive symptoms (10 cases). There were 130 benign cases: 104 adenomas, 17 adenomatous goiters, and nine atypical adenomas. There were 133 malignant cases: 94 classical papillary carcinomas, 23 follicular variants of papillary carcinoma, seven medullary carcinomas, three poorly differentiated carcinomas, three follicular carcinomas, one oncocyti (Hürthle cell) carcinoma, and two metastases.

Figure 2 Second part of the flowchart designed to score nodules with US. It defines the patterns of benign nodules. In the context of auto-immune thyroiditis, a white knight pattern is representative of numerous hypeechoic round pseudo-nodules with no halo or central vascularization. Nodular hyperplasia is representative of isoechoic confluent micronodules located within the inferior and posterior portion of one or two lobes, usually avascular and seen in simple goiters.
Assessment of imaging technique efficiency

Firstly, cytopathological results, excluding nondiagnostic lesions and lesions of indeterminate significance (atypia of undetermined significance, and suspicions of follicular neoplasm and malignancy), were used to compare the TI-RADS gray-scale score, elastography, and a combination of both methods; thus, leaving only Bethesda categories 2 and 6. This was done because the probability of mistake of these two categories is <3% compared with histopathology, in contrast to the underlying diagnostic uncertainties in case of indeterminate or nondiagnostic cytological results. By consequence, 3658 nodules were analyzed with the gray-scale score only, 991 nodules with elastography only, and 991 nodules with a combination of both methods out of a total of 4550 nodules. Secondly, the TI-RADS gray-scale score and elastography were compared with histopathological results, including cytological lesions of indeterminate significance. Sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV), and accuracy of the TI-RADS score were calculated. The TI-RADS score was considered exact when its result was benign (score 1, 2, or 3) and the final diagnosis after cytology and/or histology was benign or when its result was suspicious for malignancy (4A, 4B, or 5) and the final diagnosis was malignant.

Interobserver agreement

On a subset of 180 consecutive nodules for which patients were referred for both US examination and FNAB, each nodule was successively scored in two different rooms equipped with the same US machine by two different practitioners, the second one having no possibility of knowing the prior score. Agreement was measured by taking into account firstly the complete six-point scale and secondly by grouping scores 4A, 4B, and 5. The second approach corresponded to theoretical indications of whether to proceed to FNAB.

Estimation of the decrease in unnecessary FNABs

We considered that nodules that scored 2 and 3 could be monitored safely without FNAB unless they increase in a proven way. We used previously published data to estimate the number of nodules that were likely to grow (9, 10, 11, 12) and multiplied it by the total number of nodules scored TI-RADS 2 and 3. We considered 35% as the most realistic rate of growing nodules and chose to use it for our estimation.

Statistical analysis

Statistical analysis was performed by Medcalc (MedCalc Software, Ostend, Belgium). ROC curve analysis was used to compare the three imaging techniques and to determine the optimal cut-off value between benign and suspicious nodules. Area-under-the-curve (AUC) and P value were calculated. Differences between sensitivity and AUC for the three imaging techniques were assessed by pairwise comparison using Student’s q and Tukey tests. P values <0.05 were considered significant. Interobserver agreement was calculated using Cohen’s kappa test and Pearson’s correlation coefficient.

Approval was obtained from the Institution’s Ethic Committee. Complete oral and written information on the purpose and nature of all procedures used was given to every patient, and informed consent was also obtained.

Results

Demographic data and global results of 4550 nodules by score

The median age of patients was 54 years (range 14–85). The prevalence of nodules was lower in men (sex ratio 0.22).

Estimation of the decrease in unnecessary FNABs

We considered that nodules that scored 2 and 3 could be monitored safely without FNAB unless they increase in a proven way. We used previously published data to estimate the number of nodules that were likely to grow (9, 10, 11, 12) and multiplied it by the total number of nodules scored TI-RADS 2 and 3. We considered 35% as the most realistic rate of growing nodules and chose to use it for our estimation.
The median size of nodules was 17 mm (range 4–82). For benign nodules, the range was 5–61 mm and for malignant nodules, it was 4–82 mm. There was no nodule size threshold for indicating FNAB. A total of 252 nodules measured <10 mm.

Nodules were classified as TI-RADS 2, 3, 4A, 4B, and 5 in 4.2, 48.3, 44.5, 2.7, and 0.3% of cases respectively. Distribution of carcinomas among TI-RADS scores was 0, 4.3, 54, 32.8, and 9.3% respectively. By cytology, 92% of TI-RADS scores 2 and 3 were benign, and 1.6% were suspicious of malignancy or malignant (Bethesda categories 4, 5, and 6). TI-RADS scores 4A, 4B, and 5 represented 95.7% of carcinomas (Bethesda 6) and 91.7% of suspicious lesions (Bethesda 4 and 5).

**Diagnostic performance of gray-scale US TI-RADS score compared with cytological results (3658 cases)**

When compared with Bethesda categories 2 and 6, sensitivity, specificity, NPV, PPV, and accuracy of the gray-scale TI-RADS score were 95.7, 61, 99.7, and 62% respectively. PPV was studied for each score separately: 0% for score 2, 0.25% for score 3, 6% for score 4A, 69% for score 4B, and 100% for score 5.

**Comparison of TI-RADS score, cytological and histopathological results in the operated group**

The total number of carcinomas was 204, a rate of 4.5% (204/4550). They corresponded to 139 cases asserted by cytological analysis (among them 68 cases were confirmed by histopathology after surgery) and to 65 cases diagnosed after surgery and initially classified with the Bethesda system as category 3, 4, or 5. Results are shown in Table 3. The sensitivity of the TI-RADS score was 95.7%, specificity was 61%, NPV was 99.7%, and accuracy was 62%.

**Table 1** Comparison of the clinical efficiency of TI-RADS grey scale score alone, elastography alone and of both methods in combination with cytological results

<table>
<thead>
<tr>
<th>Imaging method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>NPV (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI-RADS gray-scale score only (3658 cases)</td>
<td>95.7</td>
<td>61</td>
<td>99.7</td>
<td>62</td>
</tr>
<tr>
<td>Elastography only (991 cases)</td>
<td>74.2</td>
<td>91.1</td>
<td>98</td>
<td>90</td>
</tr>
<tr>
<td>Combined TI-RADS (991 cases) gray-scale</td>
<td>98.5</td>
<td>44.7</td>
<td>99.8</td>
<td>48.3</td>
</tr>
<tr>
<td>elastography</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NPV, negative predictive value.

**Diagnostic performance of elastography and a combination of both methods compared with cytological results (991 cases)**

ROC curve analysis confirmed an optimal cut-off value of 0.045 between stiff and soft nodules, in agreement with a previously published study (13). Values <0.045 represented stiff nodules and vice versa. Among 1305 elastographic procedures, six were nondiagnostic because the compression curve was of poor quality. There were 49 true-positive, 861 true-negative, 17 false-negative, and 64 false-positive cases. The remaining cases represented cytologically undetermined results and were excluded. Sensitivity, specificity, NPV, PPV, and accuracy of elastography were 74.2, 91.1, 98, 37.4, and 90% respectively.

Using a combination of both methods, the gray-scale score, which was initially 3 or 4A, was raised to 4B in 85 patients, the nodule being also hard on elastography. There were 65 true-positive, one false-negative (scored 3), 414 true-negative, and 512 false-positive cases. Sensitivity, specificity, NPV, and accuracy were 98.5, 44.7, 99.8, and 48.3% respectively. Sensitivity of the combination was statistically superior to the gray-scale alone (P<0.0001), but specificity and accuracy decreased (P<0.0001). A comparison of the three methods is shown in Table 1. By ROC curve analysis, AUC was 0.868 (95% CI, 0.856–0.879; P<0.0001) for gray-scale, 0.834 (95% CI, 0.810–0.857; P<0.0001) for elastography, and 0.914 (95% CI, 0.895–0.931; P<0.0001) for a combination of both methods. The only significant difference between AUCs was between elastography alone and a combination of both methods (P<0.01). PPV was also studied for each score separately (Table 2).
gray-scale score only was 93.2% when compared with histopathological results and 95.1% (194/204) when compared with the total number of carcinomas diagnosed by cytology and histology. The sensitivity of elastography only was 41.9% and of a combination of gray-scale and elastography 96.7% (Table 3). Only one carcinoma case scored TI-RADS 3 and was soft on elastography.

**Interobserver agreement**

For the total six-point gray-scale TI-RADS score, the $\kappa$ value was 0.72 (95% CI, 0.62–0.81) and Pearson’s concordance correlation coefficient was 0.73 (95% CI, 0.66–0.79). When considering only scores 4A, 4B, and 5, which correspond to theoretical indications for biopsy according to the TI-RADS system, the $\kappa$ value was 0.74 (95% CI, 0.65–0.84) and Pearson’s coefficient was also 0.74 (95% CI, 0.67–0.80).

**Estimation of the reduction in the number of FNABs**

Nodules with TI-RADS scores of 2 or 3 represented 52.4% of all nodules. By basing the reasoning on the figure of 35% of nodules that increase in size in time, we ended in a reduction of the number of FNABs estimated at 33.8% (0.65 $Z$ scores 2 and 3 that do not grow $! 0.524$ percentage of nodules scored 2 and 3).

**Table 2** Positive predictive value of the three methods compared with cytological results

<table>
<thead>
<tr>
<th>Imaging method</th>
<th>Gray-scale only (%; 3658 cases)</th>
<th>Elastography only (%; 991 cases)</th>
<th>Combined gray-scale + elastography (%; 991 cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>4A</td>
<td>6</td>
<td>37.4</td>
<td>35.9</td>
</tr>
<tr>
<td>4B</td>
<td>69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 3** Matched results of gray-scale ti-rads score and elastography with final histopathological results – (number of cases)

<table>
<thead>
<tr>
<th>Benign histological results (130 cases)</th>
<th>Malignant histological results (133 cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bethesda’s category</strong></td>
<td><strong>Bethesda’s category</strong></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
</tr>
</tbody>
</table>

**Discussion**

Over the past two decades, widespread use of US and incidental imaging detection has contributed to an increased detection of thyroid nodules and to a threefold increase in thyroid aspirates (14). The goal to avoid unnecessary repeated US examinations and FNABs led to the development of risk stratification tools. Three other groups (1, 2, 15) have issued reports on TI-RADS, but only one group has tested the system prospectively (1).

To detect thyroid carcinomas with gray-scale US with the best possible efficiency, we used the four US signs described by Kim et al. (16) and added solid and mildly hypoechoic as a fifth sign. The four US signs reported by Kim et al. correspond to a taller-than-wide shape, irregular borders, marked hypoechogeticity, and micro-calcifications. They reported an initial sensitivity of 94%; however, in our study, they only permitted to detect 41.1% of all carcinomas. By adding a fifth sign, the sensitivity of the TI-RADS gray-scale score was 95.7% compared with cyto- and histopathological (93.2%) results. To make up with the few false negative cases, which correspond to missed carcinomas, we suggest to perform FNAB when nodules initially scored TI-RADS 3 grow in a proven way during follow-up.

On the other hand, to safely rule out malignancy, we founded our score on the published material related to asserting with US that a nodule is benign (17, 18, 19). We found a very high NPV of 99.7% for scores 2 and 3. US-TI-RADS scoring has the ability to discard malignancy with a high probability. A total of 52.4% of all nodules are classified as probably benign with a 0.3% risk of mistake.

The sensitivity of elastography was 74.2% compared with cyto- and histopathological (41.9%) results. Specificity was 91.1 and 86.4% respectively, and accuracy 90 and 68%. A hard nodule should always be thought of as highly suspicious of malignancy but low stiffness is frequently encountered in carcinomas. In our study, the sensitivity of elastography for the detection of thyroid carcinoma was lower than expected according to the meta-analysis by Bojunga et al. (20).
Most of our cases were referred for FNAB of a single nodule, therefore constituting a potential selection bias. This discrepancy in results may also be explained by the fact that the study by Bojunga et al. was largely based on qualitative elastographic techniques, different from the quantitative technique we used. In agreement with our results, Unütürk et al. (21) noted a lower sensitivity and specificity of US elastography for the diagnosis of malignant thyroid nodules than previously reported.

The combination of gray-scale US and elastography can be used in two separate ways. Firstly, it can increase sensitivity because it can detect a small percentage of carcinomas omitted by gray-scale alone. Trimboli et al. (22) used signs for detecting carcinomas that were identical to our group and reported a sensitivity of only 85% with gray-scale. This value increased to 97% when gray-scale was combined with elastography. However, Moon et al. (23) found that the combination of both methods showed inferior performance in the differentiation of malignant and benign thyroid nodules compared with gray-scale US features. Secondly, a combination of gray-scale US and elastography can raise specificity. Biopsying or monitoring nodules scored TI-RADS 4A with gray-scale that are soft on elastography could be discussed, since the risk of carcinoma rises. Biopsying or monitoring nodules scored TI-RADS 4A with gray-scale that are soft on elastography can be used in two separate ways. Firstly, it can increase sensitivity because it can detect a small percentage of carcinomas omitted by gray-scale alone. Trimboli et al. (22) used signs for detecting carcinomas that were identical to our group and reported a sensitivity of only 85% with gray-scale. This value increased to 97% when gray-scale was combined with elastography. However, Moon et al. (23) found that the combination of both methods showed inferior performance in the differentiation of malignant and benign thyroid nodules compared with gray-scale US features. Secondly, a combination of gray-scale US and elastography can raise specificity. Biopsying or monitoring nodules scored TI-RADS 4A with gray-scale that are soft on elastography could be discussed, since the risk of carcinoma dropped to 1.6% in our study vs 6% among all nodules that scored 4A (Table 2).

**Enhance interobserver agreement**

While US scanning is noninvasive, a major concern is that it is operator dependent. Regarding reproducibility, Hambly et al. (24) studied the interobserver agreement of a five-point malignancy rating scale very similar to the TI-RADS score on 101 nodules. The interobserver agreement on whether to proceed to biopsy was fair to substantial, ranging from a $\kappa$ value of 0.38–0.69. In our study on 180 nodules, for the complete six-point scale, corresponding to the likelihood for malignancy, $\kappa$ value and Pearson’s concordance correlation coefficient were respectively 0.72 and 0.73. Considering that nodules with scores of 2 and 3 should not undergo biopsy and that scores of 4 and 5 should, the interobserver agreement for biopsy had a $\kappa$ value of 0.74. In both cases, interobserver agreement was substantial.

**Rationalize the indications for FNABs and avoid unnecessary ones**

Proven high sensitivity, NPV, and interobserver agreement could allow to monitor the nodules with a TI-RADS score of 2 or 3, which represent 52.4% of all nodules, unless they increase in volume in a proven way. Considering that ~35% of thyroid nodules increase in size, we ended in a reduction of the number of FNABs estimated at 33.8%. A reduction in unnecessary FNABs obtained by using US scanning risk stratification has been studied in three reports. In a series of 450 nonpalpable nodules, Leenhardt et al. (25) reported that indication of US-FNAB appears judicious in centimetric or supracentimetric nodules or in solid and hypoechoic ones. This disease management would avoid 16% of unnecessary biopsies. In 2006, Cappelli et al. (5) performed a retrospective study on 6135 nodules. These authors suggested performing FNAB on nodules with the taller-than-wide sign or on those having at least two of the following three signs: microcalcifications, ill-defined borders, and hypoechogenicity. This strategy would omit 0.9% of cancers, while reducing the number of FNABs by 28%. Finally, Horvath et al. in 2009 (1) stated that the indication of FNABs was useless in 34% of all cases. Among these, only 1% corresponded to carcinomas according to the cytological results.

The results of these previous reports are consistent with ours. In our practice, scoring is now used daily to select nodules for US-FNAB. Based on the French Endocrine Society consensus (26) and its indications for FNAB for nodules >7 mm, we suggest taking into account the importance of TI-RADS scoring and propose indications of FNAB as follows: i) between 7–10 mm, we recommend FNAB for scores of 4B and 5 or when the patient has risk factors (family history of thyroid cancer, neck irradiation during childhood) or if there is focal PET–SCAN uptake and ii) >10 mm, we recommend FNAB for a score of 4A or 3 that grows in a proven way (26), that is to say 2 mm or more in two perpendicular planes and more than 20% in volume. Of course, these are general guidelines, and from a medical and legal point of view, the indication of FNAB should always be tailored to each patient.

This study has some limitations. Firstly, several US signs are not taken into account in the flowchart, including macrocalcifications, the halo sign, and central vascularization. This was done to simplify the system and to increase interobserver agreement; however, these signs and others may be instrumental in refining one’s judgment on a particular nodule. This is the counterpart of the process of systematization in that it truncates the complexity of reality. Secondly, clinical behavior, TSH assessment, and results of scintigraphy, when available, should also be taken into account for the indication of FNAB. Thirdly, the fact that very small nodules were subjected to FNAB could be disputable; however, the aim of our study was to test the TI-RADS score as independently as possible of those factors. Fourthly, this study represents a single specialized thyroid clinic’s work; thus, results have to be confirmed with multicenter studies and tested by nonspecialized US practitioners. Fifthly, the decrease in the number of unnecessary biopsies was only estimated and needs a proper randomized prospective trial for itself. Finally, histological confirmation is unavailable for 53.6% of cytological cases categorized Bethesda 4 to 6. Part of this bias is compensated for by the size of the population included in this series and its prospective technique.
In conclusion, the TI-RADS score is a six-point scale for risk stratification of thyroid nodules. It gives us the ability to detect most thyroid carcinomas and to assert benign status with good reliability for more than 50% of all nodules. Elastography can be used to raise sensitivity or specificity. Interobserver agreement is substantial. TI-RADS could lead to a significant decrease of the number of unnecessary FNABs.

Declaration of interest
The 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.

References