Endocrine (2019) 65:121–131
https://doi.org/10.1007/s12020-019-01886-0

ORIGINAL ARTICLE

Check for updates

Comparison and preliminary discussion of the reasons for the differences in diagnostic performance and unnecessary FNA biopsies between the ACR TIRADS and 2015 ATA guidelines

Xiao-Li Wu[1] · Jia-Rui Du[2] · Hui Wang[2] · Chun-Xiang Jin[2] · Guo-Qing Sui[2] · Dong-Yan Yang[2] · Yuan-Qiang Lin[2] · Qiang Luo[2] · Ping Fu[2] · He-Qun Li[2] · Deng-Ke Teng[2]

Received: 10 December 2018 / Accepted: 25 February 2019 / Published online: 4 March 2019

Abstract

Objectives (1) To compare the American College of Radiology (ACR) thyroid imaging reporting and data system (TIRADS) and American Thyroid Association (ATA) guidelines for thyroid nodules with regard to diagnostic performance and effectiveness at reducing the number of fine-needle aspiration (FNA) biopsies and to preliminarily discuss the reasons for the differences and (2) to compare the diagnostic performance of the two guidelines in the subgroup of nodules <1 cm in diameter.

Materials and methods In the present study, 1000 thyroid nodules in 894 consecutive patients with final diagnoses were included; these thyroid nodules were investigated via FNA biopsies in our hospital. The ultrasound (US) features of the thyroid nodules were reviewed and stratified according to the categories defined by the ACR TIRADS and ATA guidelines.

Results Compared with the ACR TIRADS guidelines, the ATA guidelines had a higher sensitivity (93.4% ($P < 0.001$)) and a larger negative predictive value (NPV) (85.3% ($P = 0.034$)). Compared with the ATA guidelines, the ACR TIRADS guidelines had a higher specificity (66.0% ($P < 0.001$)), a greater PPV (73.6% ($P = 0.001$)), and greater accuracy (75.5% ($P = 0.017$)). Compared with the ATA guidelines, the ACR TIRADS guidelines resulted in significantly fewer unnecessary FNA biopsies ($P = 0.007$).

Conclusions This study suggests that both the ACR TIRADS and ATA guidelines have unique strengths with regard to their diagnostic performance. In terms of reducing the number of FNA biopsies, the ACR TIRADS guidelines were superior to the ATA guidelines.

Keywords Thyroid nodule · Thyroid cancer · Ultrasound diagnosis

Abbreviations

ACR	American College of Radiology
ATA	American Thyroid Association
FNA	fine-needle aspiration
CNB	core-needle biopsy
TIRADS	thyroid imaging reporting and data system
US	ultrasound

Introduction

The incidence of thyroid cancer has increased substantially in the past few decades [1–3], although it has stabilized in recent years [4]. The accurate diagnosis and management of thyroid nodules are now considered to be of great importance [5]. Various guidelines have recommended that ultrasound (US) be performed as the first-line examination for the diagnosis of thyroid nodules [6–10].

According to the 2015 American Thyroid Association (ATA) management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer, thyroid nodules are stratified into five types based on their US characteristics [10]. Each type has a different risk of malignancy, and a fine-needle aspiration (FNA) biopsy may be recommended due to a nodule's size. Some studies have

These authors contributed equally: Xiao-Li Wu, Jia-Rui Du

✉ Deng-Ke Teng
tengdke@163.com

[1] Department of Ultrasound, The First Hospital of Jilin University, Changchun 130000 Jilin, China

[2] Department of Ultrasound, China-Japan Union Hospital of Jilin University, Changchun 130000 Jilin, China

sought to compare the ATA guidelines with other guidelines [11–13].

In 2017, the thyroid imaging reporting and data system (TIRADS) was described by the American College of Radiology (ACR) to guide the differential diagnosis and decisions regarding the need for FNA biopsies of thyroid nodules [14]. In the ACR TIRADS, thyroid nodules are stratified into five types according to their US scores. FNA biopsies may then be recommended based on a nodule's ACR TIRADS type and its largest diameter. Some scholars have published studies on the use of this classification system [15–25]. A few of those studies make comparisons between the ACR TIRADS and the ATA guidelines with regard to the efficacy of the management of thyroid nodules [17, 18, 21–23]. However, none of these studies discussed the probable reasons for the difference in diagnostic performance between the two sets of guidelines; such a discussion would be useful for clinicians, who seek to provide optimal recommendations to patients with thyroid nodules according to the actual condition of the patient. Moreover, only Ha EJ's study referred to the diagnostic performance of the two sets of guidelines for nodules <1 cm in diameter, and investigating that performance was not their main research aim [18]. Though in most cases, the analysis of thyroid nodules larger than 1 cm can be satisfactory, the accurate qualitative analysis of nodules <1 cm is also needed for some patients with high-risk factors (such as childhood exposure history, familial thyroid cancer, lymphadenopathy, rapid nodule growth, and hoarseness).

Accordingly, the aims of this study were as follows: (1) to compare the ACR TIRADS and ATA guidelines for thyroid nodules with regard to their diagnostic performance and effectiveness at reducing the number of FNA biopsies and to preliminarily discuss the reasons for the differences and (2) to compare the diagnostic performance of the two guidelines in the subgroup of nodules <1 cm in diameter.

Materials and methods

This study was approved by the Ethics Committee of the China-Japan Union Hospital of Jilin University, Changchun, China.

Patients

From April 2016 to March 2017, FNA biopsies were performed on 1315 consecutive nodules in our department. Among these nodules, 315 nodules were excluded since a final diagnosis was not obtained. Among those excluded nodules, 98 had nondiagnostic or unsatisfactory FNA biopsy results, and 217 nodules had indeterminate cytology findings atypia of undetermined significance/follicular lesion of undetermined significance (AUS/FLUS), follicular neoplasm (FN), suspicious for malignancy (SUS). In total, 1000 thyroid nodules (the number was prespecified by a power analysis) with final diagnoses in 894 consecutive patients were included in this study (Fig. 1).

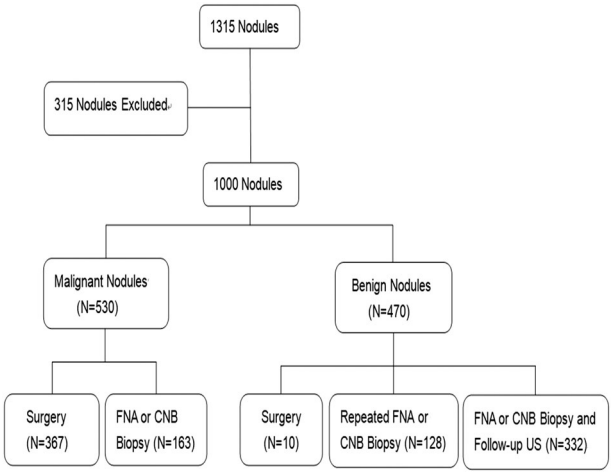

Fig. 1 Flowchart shows the participants of our study

With regard to the malignant nodules ($n = 530$), 387 nodules removed during surgery were diagnosed based on the histopathologic results after surgery, 163 nodules that were not excised were diagnosed by FNA biopsy or core-needle biopsy (CNB). With regard to the benign nodules ($n = 470$), 10 nodules removed during surgery were diagnosed by the histopathologic results after surgery, 128 nodules were diagnosed by at least two FNA biopsies or CNBs, and 332 nodules were diagnosed by FNA biopsy or CNB with at least 12 months of follow-up during which the size of the nodule decreased or remained stable.

US examination and image analysis

All US examinations were performed with a 4–9 MHz or 9–14 MHz linear probe (MINDRAY® DC-8Exp color Doppler Ultrasound System, Shenzhen, China). The images and videos of each nodule were recorded by a physician with over 20 years of experience in US examinations. Most nodules were detected by a 9–14 MHz linear probe for information acquisition. However, when a nodule is large, and the position of the nodule is deep, the high-frequency linear array probe sometimes cannot display the nodule very well. In those instances, a 4–9 MHz linear probe, which can increase the penetration of the ultrasonic waves, should be used to detect the deep portion of the nodule. All nodules were jointly evaluated by three doctors with over 15 years of experience in analyzing thyroid USs. Without any previous information about the pathologic results and basing their analyses on the ACR and ATA guidelines, the doctors

evaluated the composition (cystic, solid, mixed cystic, and solid), echogenicity (anechoic, hyperechoic, isoechoic, hypoechoic, very hypoechoic), shape (taller-than-wide or not), margins (smooth or ill-defined, lobulated or irregular, extrathyroidal extension), and echogenic foci (macrocalcifications, peripheral calcifications, and punctuate echogenic foci). All the evaluated results were obtained by consensus among the three doctors. Instead of first analyzing all the nodules and then discussing their results to arrive at a consensus, the three doctors analyzed each nodule and immediately discussed that nodule to reach a consensus before moving on to analyzing the next nodule. Disagreements among the doctors were resolved by selecting the conclusion reached by the majority. After a consensus was achieved for a nodule, the agreement was used as a standard for the subsequent analyses. After reviewing the nodules, the doctors scored and categorized the nodules according to the ACR TIRADS guidelines, e.g., TR1 (0 point), TR2 (2 points), TR3 (3 points), TR4 (4–6 points), and TR5 (≥7 points), and according to the ATA guidelines, e.g., benign, very low suspicion, low suspicion, intermediate suspicion, and high suspicion.

Statistical analyses

SPSS 23.0 for Windows was employed for the statistical analyses. The receiver operating characteristic (ROC) curve was generated based on the pathological results of all 1000 thyroid nodules to determine the cutoff values for benign and malignant thyroid nodules according to the ACR and ATA guidelines. With regard to the diagnosis of the thyroid nodules, we calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of the two sets of guidelines and compared them with the McNemar test. Based on the pathological results, we calculated the malignancy risk of the ACR scores and categories and the ATA risk stratification grades. A similar analysis was performed with the subgroup of nodules ≤1 cm in diameter. Furthermore, for nodules larger than 1 cm, the rate of unnecessary FNA biopsies was compared following the criteria in both the ACR and ATA guidelines. A significant difference was defined as $P < 0.05$.

Results

General characteristics

In the present study, 1000 thyroid nodules with final diagnoses in 894 consecutive patients were finally included. In total, 794 patients had one nodule each, 95 patients had two nodules, 4 patients had three nodules, and 1 patient had four nodules. Of the 1000 thyroid nodules, 470 (47%) were diagnosed as benign, and 530 (53%) were diagnosed as malignant. There were 336 lesions <1 cm in diameter. The numbers of nodules in each size category are listed in Table 1.

Table 1 Number of nodules in each size category in the group

Size (cm)	No. of nodules	Benign nodules	Malignant nodules	Mean ± SD
<1.0	336 (33.6)	165 (16.5)	171 (17.1)	0.81 ± 0.13
1.0–1.5	504 (50.4)	226 (22.6)	278 (27.8)	1.19 ± 0.14
1.5–2.0	77 (7.7)	43 (4.3)	34 (3.4)	1.68 ± 0.13
2.0-2.5	33 (3.3)	15 (1.5)	18 (1.8)	2.17 ± 0.15
≥2.5	50 (5.0)	21 (2.1)	29 (2.9)	3.89 ± 1.01

Data in parentheses are percentages

SD standard deviation

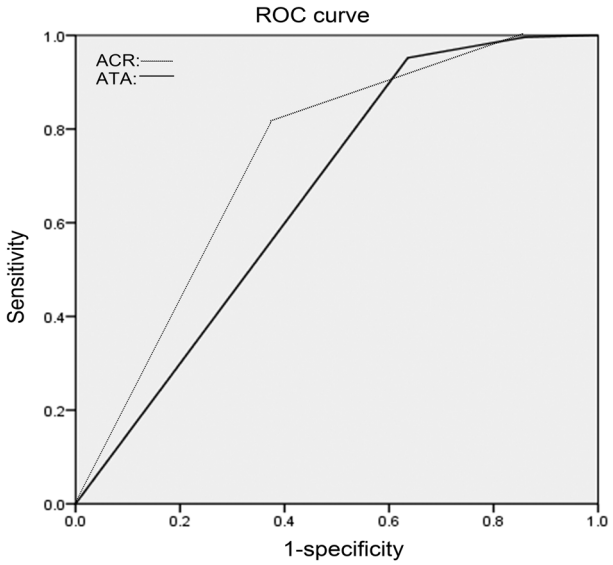

Fig. 2 The ROC curve of the ACR and ATA guideline

Comparison of diagnostic performance between the ATA and ACR guidelines

All 1000 thyroid nodules could be classified following the ACR guidelines, and 940 could be classified according to the ATA guidelines. In total, 6% (60 of 1000) of the nodules did not meet the criteria for any pattern in the ATA guidelines; 10 of the 60 nodules were malignant, yielding a malignancy risk of 16.7% (10 of 60). As suggested by the ROC curve (Fig. 2), the ACR guidelines could distinguish between benign and malignant nodules with a cutoff value of 4.5, meaning that when the TR (TIRADS) is 5, the diagnosis is that the nodule is malignant, and when the TR

Table 2 Comparison of diagnostic performance for malignant thyroid nodules in various guidelines

Guideline	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Accuracy (%)
All					
ACR	84.0 (445/530) [80.6, 86.9]	66.0 (310/470) [61.6, 70.1]	73.6 (445/605) [69.9, 76.9]	78.5 (310/394) [74.4, 82.5]	75.5 (755/1000) [72.7, 78.1]
ATA	93.4 (495/530) [90.9, 95.2]	43.2 (203/470) [38.8, 47.7]	65.0 (495/762) [61.5, 68.3]	85.3 (203/238) [80.2, 89.3]	69.8 (698/1000) [66.9, 72.6]
P-value	0.000	0.000	0.001	0.034	0.017
<1.0 cm					
ACR	77.8 (133/171) [71.0, 83.4]	70.3 (116/165) [62.9, 76.8]	73.1 (133/182) [66.2, 79.0]	75.3 (116/154) [67.9, 81.5]	74.1 (249/336) [69.2, 78.5]
ATA	93.6 (160/171) [88.7, 96.5]	43.0 (71/165) [35.7, 50.7]	63.0 (160/254) [56.9, 68.7]	86.6 (71/82) [77.4, 92.5]	68.8 (231/336) [63.6, 73.5]
P-value	0.000	0.000	0.027	0.042	0.220

Data in parentheses are the raw data used to calculate the percentages, and data in brackets are 95% confidence intervals

P-value is calculated by McNemar test

ACR American College of Radiology, *ATA* 2015 American Thyroid Association Management Guidelines for adult patients with thyroid nodules and differentiated thyroid cancer

is 4 or less, the diagnosis is that the nodule is benign; this cutoff value yielded the most accurate diagnoses based on the ACR guidelines. The ATA guidelines could differentiate between benign and malignant nodules with a cutoff value of 4.5. In other words, when the risk stratification is high suspicion, the diagnosis is that the nodule is malignant, and when the risk stratification is intermediate suspicion and below, the diagnosis is that the nodule is benign; this cutoff value led to the best diagnostic accuracy for the ATA guidelines.

The diagnostic performance of the ACR and ATA guidelines were evaluated based on the above criteria. Compared with the ACR guidelines, the ATA guidelines were more sensitive (93.4% ($P < 0.001$)) and had a larger NPV (85.3% ($P = 0.034$)). Compared with the ATA guidelines, the ACR guidelines were more specific (66.0% ($P < 0.001$)), had a larger PPV (73.6% ($P = 0.001$)), and were more accurate (75.5% ($P = 0.017$)). The comparison of the diagnostic performance between the two guidelines is shown in Table 2.

The malignancy risk of different categories in the ACR and ATA guidelines

In this study, the malignancy risk of the ATA high-risk classification is 65.0%, which is close to the theoretical value of 70–90%. The malignancy risk of the ACR TR5 category is 73.6%, which is much higher than the theoretical 20%. Table 3 shows the malignancy risk for each category in the two guidelines. Moreover, as the ACR score increases, so does the malignancy risk. In addition, based on the pathological results, we analyzed the malignancy risks of various malignant signs in line with the malignant signs on US proposed in the ACR guidelines. Among them, the malignancy risks of very hypoechoic, taller-than-wide and punctuate echogenic nodules were all above 70% (Table 4).

Comparison of the rates of unnecessary FNA biopsies

For 664 nodules larger than 1 cm (305 benign and 359 malignant), the rates of unnecessary FNA biopsies were 23.3% when following the ACR guidelines and 35.2% when following the ATA guidelines. The difference is significant ($P = 0.011$). Furthermore, 327 out of 359 malignant lesions (91.1%) were detected by following the ACR FNA biopsy guidelines, and 348 out of 359 malignant lesions (96.9%) were detected by following the ATA FNA biopsy guidelines. Though the slightly more malignant lesions were detected by following the ATA guidelines compared with the ACR guidelines, the difference was not significant ($P = 0.180$). The comparison of the rates of

Table 3 Comparison of malignancy risk according to category in various guidelines

Guideline	Benign nodules		Malignant nodules		Total		Suggested risk of malignancy (%)	Calculated risk of malignancy (%)	
	All ($n = 470$)	<1 cm ($n = 165$)	All ($n = 530$)	<1 cm ($n = 171$)	All ($n = 1000$)	<1 cm ($n = 336$)		All	<1 cm
ACR 2017									
Highly suspicious	160	49	445	133	605	182	>20	73.6	73.1
Moderately suspicious	261	100	83	38	344	138	5–20	24.1	27.5
Mildly suspicious	34	11	2	0	36	11	5	5.6	0
Not suspicious	15	5	0	0	15	5	<2	0	0
Benign	0	0	0	0	0	0	<2	0	0
ATA 2015									
High suspicion	267	94	495	160	762	254	>70–90	65.0	63.0
Intermediate suspicion	94	48	23	11	117	59	10–20	19.7	18.6
Low suspicion	54	11	2	0	56	11	5–10	3.6	0
Very low suspicion	5	5	0	0	5	5	<3	0	0
Benign	0	0	0	0	0	0	<1	0	0
Not specified[a]	50	7	10	0	60	7	…	16.7	0

All acronyms are as given in Table 2

[a]Isoechoic nodules with suspicious US features (irregular margins, microcalcifications, and taller-than-wide shape) could not be classified as any specific pattern with the ATA criteria

unnecessary FNA biopsies between the two guidelines is shown in Table 5.

Comparison of diagnostic performance between the two guidelines for nodules <1 cm in diameter

In this study, there were 336 nodules <1 cm in diameter. In addition, the diagnostic performance of the ATA and ACR guidelines for those nodules was different (Table 2). Compared with the ACR guidelines, the ATA guidelines had a higher sensitivity (93.6% ($P < 0.001$)) and a larger NPV (86.6% ($P = 0.042$)). Compared with the ATA guidelines, the ACR guidelines had a higher specificity (70.3% ($P < 0.001$)) and a larger PPV (73.1% ($P = 0.027$)). The accuracy of the ACR guidelines was slightly higher (74.1%) than that of the ATA guidelines, although the difference was not significant ($P = 0.220$). As the risk stratification level increases, the malignancy risk of nodules <1 cm in diameter also increases. The malignancy risk of the ATA high-risk classification is 63.0%, which is close to the theoretical value of 70–90%. The malignancy risk of ACR TR5 category is 73.1%, which is higher than the theoretical 20% (Table 3). As the ACR score increases, the malignancy risk of nodules <1 cm in diameter also increases. We also analyzed the malignancy risks of various malignant signs in line with the US malignant signs proposed in the ACR guidelines (Table 4). Among these signs, the malignancy risks of very hypoechoic and punctuate echogenic nodules are both above 70%.

Discussion

This study suggests that ACR and ATA guidelines had different strengths with regard to diagnosing thyroid nodules. Compared with the ACR guidelines, the ATA guidelines had a higher diagnostic sensitivity and a larger NPV. However, compared with the ATA guidelines, the ACR guidelines yielded a higher specificity, larger PPV, and greater accuracy. Compared with the ATA guidelines, the ACR guidelines had a lower rate of unnecessary FNA biopsies ($P = 0.007$). For the subgroup of nodules <1 cm in diameter, compared with the ACR guidelines, the ATA guidelines still had a higher sensitivity and a larger NPV, while comparing with the ATA guidelines, the ACR guidelines still had a higher specificity and larger PPV. However, the diagnostic accuracies of the two guidelines for nodules <1 cm were not significantly different.

Diagnostic performance

In previous studies [17, 18, 21, 23], the diagnostic performance of the ATA and ACR guidelines have been compared for thyroid nodules larger than 1 cm. However, only Ha EJ's study referred to the diagnostic performance for nodules <1 cm in the ACR and ATA guidelines, and it was not their main research aim [18]. Moreover, none of these studies discussed the probable reasons for the difference in performance of the two guidelines in diagnosing thyroid nodules.

Table 4 Frequency of US findings in all (1000 nodules) and <1 cm (336 nodules) thyroid nodules based on expert consensus

Findings	Nodules		Benign		Malignant		Malignant rate (%)	
	All ($n = 1000$)	<1 cm ($n = 336$)	All ($n = 470$)	<1 cm ($n = 165$)	All ($n = 530$)	<1 cm ($n = 171$)	All	<1 cm
Composition								
Cystic or spongiform	1 (0.1)	0	1 (0.2)	0	0	0	0	0
Cystic and solid	52 (5.2)	13 (3.9)	38 (8.1)	11 (6.7)	14 (2.6)	2 (1.2)	26.9	26.9
Solid	947 (94.7)	323 (96.1)	431 (91.7)	154 (93.3)	516 (97.4)	169 (98.8)	54.5	52.3
Echogenicity								
Anechoic	0	0	0	0	0	0	…	…
Hyperechoic or isoechoic	117 (11.7)	21 (6.3)	97 (20.6)	18 (10.9)	20 (3.8)	3 (1.8)	17.1	14.3
Hypoechoic	778 (77.8)	272 (81.0)	343 (73.0)	136 (82.4)	435 (82.1)	136 (79.5)	55.9	50.0
Very hypoechoic	105 (10.5)	43 (12.8)	30 (6.4)	11 (6.7)	75 (14.2)	32 (18.7)	71.4	74.4
Shape								
Taller-than-wide	374 (37.4)	158 (47.0)	109 (23.2)	54 (32.7)	265 (50.0)	104 (60.8)	70.9	65.8
Not taller-than-wide	626 (62.6)	178 (53.0)	361 (76.8)	111 (67.3)	265 (50.0)	67 (39.2)	42.3	42.3
Margi								
Smooth or ill defined	358 (35.8)	155 (46.1)	244 (51.9)	92 (55.8)	114 (21.5)	63 (36.8)	31.8	40.6
Irregular or lobulated	611 (61.1)	181 (53.9)	216 (46.0)	73 (44.2)	395 (74.5)	108 (63.2)	64.6	59.7
Extrathyroidal extension	31 (3.1)	0	10 (2.1)	0	21 (4.0)	0	67.7	0
Echogenic foci[a]								
Macrocalcifications	55 (5.5)	13 (3.9)	35 (7.4)	10 (6.1)	20 (3.8)	3 (1.8)	36.4	23.1
Peripheral calcifications	32 (3.2)	11 (3.3)	21 (4.5)	6 (3.6)	11 (2.1)	5 (2.9)	34.4	45.5
Peripheral and macrocalcifications	30 (3.0)	6 (1.2)	15 (3.2)	5 (3.0)	15 (2.8)	1 (0.6)	50.0	16.7
Punctate	264 (26.4)	35 (10.4)	75 (16.0)	6 (3.6)	189 (35.7)	29 (17.0)	71.6	82.9
Punctate and macrocalcifications	34 (3.4)	2 (0.6)	10 (2.1)	0	24 (4.5)	2 (1.2)	70.6	100.0
Punctate and peripheral calcifications	5 (0.5)	13 (3.9)	0	0	5 (0.9)	2 (1.2)	100.0	100.0

Unless otherwise indicated, data are number of nodules, and data in parentheses are percentages

[a]Nodules could have more than one type of echogenic foci

In this study, thyroid nodules including <1 cm were analyzed, and the cutoff value of the optimal diagnostic performance was obtained from the ROC curve. The high suspicion nodules of the ATA guidelines and the TR5 nodules of the ACR guidelines served as criteria for the definition of malignancy. Analysis suggested that compared with the ATA guidelines, the ACR guidelines have a higher specificity (66.0%, $P < 0.001$), have a larger PPV (73.6%, $P = 0.001$), and are more accurate (75.5%, $P = 0.017$), while comparing with the ACR guidelines, the ATA guidelines have a higher sensitivity (93.4%, $P < 0.001$) and a larger NPV (85.3%, $P = 0.034$). We thought this was because the ATA guidelines use lobes or infiltration at the edge of the nodule as a critical grading indicator. For instance, when a solid hypoechoic thyroid nodule with edge lobulation or marginal infiltration is stratified as a highly suspicious nodule, this nodule will be defined as malignant. The same nodule in the ACR guideline is classified as solid (2 points), hypoechoic (2 points), and having edge irregularity (2 points); this nodule receives a total score of 6 points and is classified as TR4, while it indicates that it is a benign nodule. If the pathological diagnosis of this nodule is benign (Fig. 3a), the ACR guidelines will have diagnosed it correctly and will have a higher specificity. If the pathological diagnosis of this nodule is malignant (Fig. 3b), the ATA guidelines will have diagnosed it correctly and will have a higher sensitivity.

However, the ACR guidelines could have a greater accuracy because unlike the ATA guidelines, they take into account macrocalcification. For instance, for a solid, hypoechoic, marginal irregular nodule with a pathological diagnosis of malignancy, the ATA guidelines rate the

Table 5 Comparison of unnecessary fine-needle aspiration biopsy rates for diagnosis of thyroid cancer in all nodules

Guideline	No. of FNABs	No. of malignant nodules among FNAB nodules	No. of benign nodules among FNAB nodules (A)	Unnecessary FNAB rate (A/664[a])	False-positive rate[b] (A/470[c])	Positive rate (%)	Detection rate (%)	*P*-value
ACR	482	327 (67.8)	155 (32.2)	23.3 [20.3, 26.7]	33.0 [28.9, 37.4]	67.8 [63.5, 71.9]	49.2 [45.5, 53.0]	0.007
Highly suspicious	423	312 (73.8)	111 (26.2)	16.7 [14.1, 19.8]	23.6 [20.0, 27.7]	73.8 [69.4, 77.7]	47.0 [43.2, 50.8]	0.010
Moderately suspicious	53	14 (26.4)	39 (73.6)	5.9 [4.3, 8.0]	8.3 [6.1, 11.2]	26.4 [16.3, 39.7]	2.1 [1.2, 3.5]	0.477
Mildly suspicious	6	1 (1.7)	5 (83.3)	0.8 [0.3, 1.8]	1.1 [0.4, 2.5]	16.7 [1.1, 5.8]	0.2 [0.0, 0.9]	0.449
ATA	582	348 (59.8)	234 (40.2)	35.2 [31.7, 39.0]	49.8 [45.3, 54.3]	59.8 [55.8, 63.7]	52.4 [48.6, 56.2]	0.007
High suspicion	508	335 (65.9)	173 (34.1)	26.1 [22.9, 29.5]	36.8 [32.6, 41.3]	65.9 [61.7, 69.9]	50.5 [46.7, 54.2]	0.010
Intermediate suspicion	58	12 (20.7)	46 (79.3)	7.0 [5.2, 9.1]	9.8 [7.4, 12.8]	20.7 [12.1, 32.9]	1.8 [1.0, 3.2]	0.477
Low suspicion	16	1 (6.3)	15 (93.8)	2.3 [1.3, 3.7]	3.2 [1.9, 5.2]	6.3 [0.2, 30.2]	0.2 [0.0, 0.9]	0.449
P -value	...	...	...	0.011	0.001	0.007	0.434	...

Unless otherwise indicated, data in parentheses are percentages. Data in brackets are 95% confidence intervals

P-value is calculated by McNemar test

FNAB fine-needle aspiration biopsy; for all other acronyms see Table 2

[a]664 are to exclude 336 nodules <1 cm in 1000 nodules

[b]Data in parentheses are the raw data from which the false-positive rate was calculated

[c]470 are benign nodules in 1000 nodules

nodule as highly suspicious, which is a correct diagnosis, whereas the ACR guidelines rate the nodule 6 points (TR4), which is a misdiagnosis. However, when this nodule is accompanied by macrocalcification, the ACR guidelines would rate the nodule 7 points (TR5), which is a correct diagnosis (Fig. 3c). Furthermore, the ACR guidelines divide hypoechoic into hypoechoic and very hypoechoic. Very hypoechoic is defined as decreased echogenicity relative to that of the adjacent neck musculature [26], and ACR guidelines increase the score for very hypoechoic to 3 points, which could also be a reason for their superior diagnostic accuracy. For instance, the score of a malignant, solid, very hypoechoic, irregular edged thyroid nodule (Fig. 3d) would be 7 points (TR5) according to the ACR guidelines, which would be the correct diagnosis. In the present study, the malignancy rate of very hypoechoic nodules was 71.4%, which was much higher than the 55.9% of hypoechoic nodules (Table 4). This is evidence supporting the supposition that very hypoechoic nodules are more likely to be malignant. The diagnostic accuracy of the ACR guidelines (75.5%) is higher than that of the ATA guidelines (69.8%, $P = 0.004$), which is probably caused by the scoring of macrocalcification and very hypoechoic nodules in the ACR guidelines. At the same time, we should highlight that the ROC-derived cutoff value is not recommended as a diagnostic standard to be used in clinical practice as it may underestimate the real accuracy.

In addition, in previous studies, it was found that nearly 3.4–13.9% of lesions could not be evaluated by the ATA guidelines [22, 23, 27]. In this study, 60 of the 1000 lesions (6%) were not evaluated by the ATA guidelines (Fig. 3e) since the stratification in the ATA guidelines did not include the evaluation of hyperechoic nodules or isoechoic nodules with malignant signs (taller-than-wide, punctate, marginal irregularities, etc.). However, 10 (16.7%) of these nodules were malignant. This may be a disadvantage of the ATA guidelines, which should be investigated in the future.

FNA performance

Our study suggests that following the ACR guidelines significantly reduce the number of unnecessary FNA biopsies compared with following the ATA guidelines. Some previous studies suggested that the ACR guidelines were more effective at guiding decisions regarding performing FNA biopsies [16–18, 22, 23], whereas some researchers believe that the FNA biopsy threshold of the ACR guidelines needs more consideration and prospective validation [15]. In this study, the number of FNA biopsies performed when following the ACR guidelines was 482, of which 327 were malignant and 155 were benign (23.3%); these numbers are significantly lower than the 582 FNA biopsies

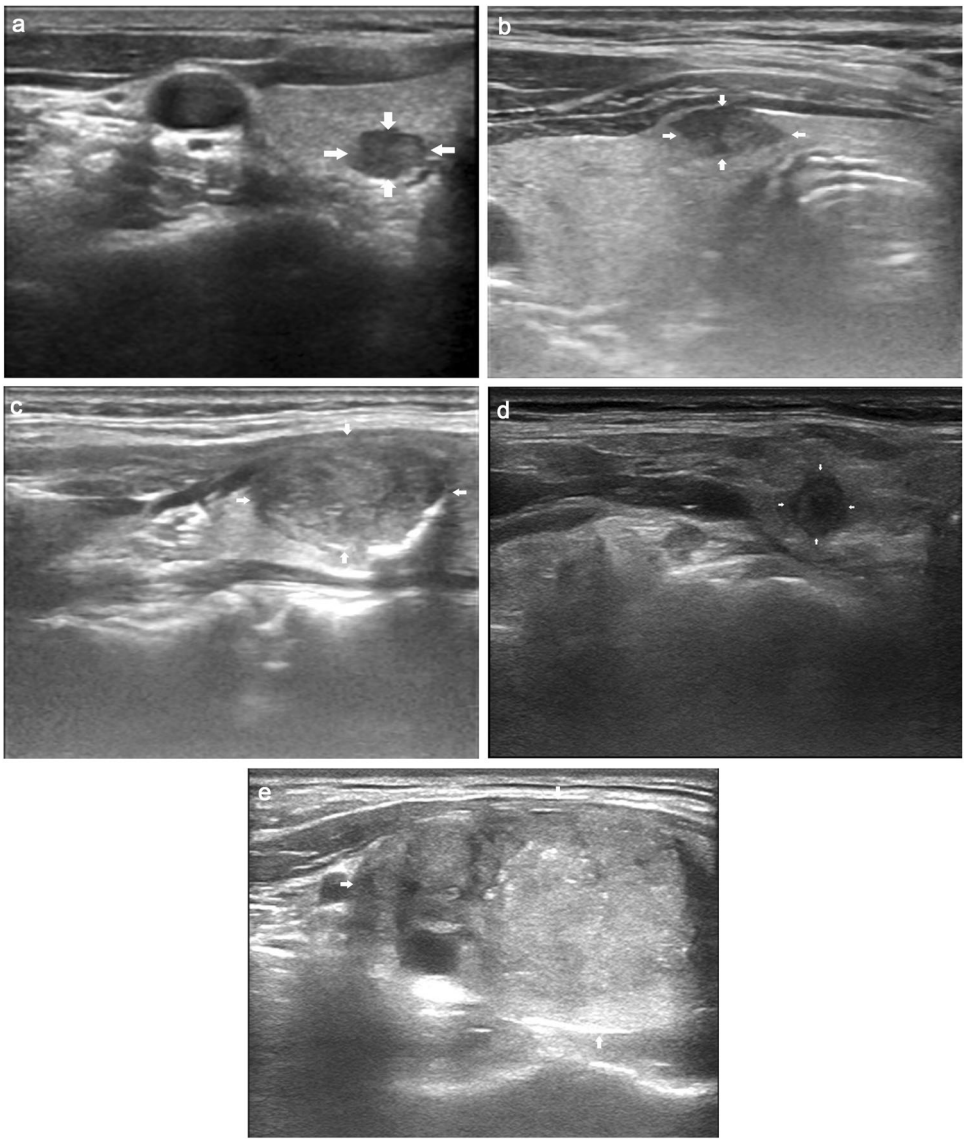

Fig. 3 **a** This is a solid, hypoechoic, edge-leaved nodule. The pathological diagnosis is a benign nodule. The ACR guideline is classified as TR4 and is a benign nodule. The ATA guideline is classified as highly suspicious and is a malignant nodule. The ACR guidelines are diagnosed correctly, indicating that the ACR guidelines have a higher specificity. **b** This is a solid, hypoechoic, edge irregularly nodule. The pathological diagnosis is a malignant nodule. The ACR guideline is classified as TR4 and is a benign nodule. The ATA guideline is classified as highly suspicious and is a malignant nodule. The ATA guidelines are diagnosed correctly, indicating that the ATA guidelines have a higher sensitivity. **c** This is a solid, hypoechoic, macrocalcification, irregularly shaped nodule. The pathological diagnosis is a malignant nodule. The ATA guideline is classified as highly suspicious and is a malignant nodule. The ACR guidelines are classified as solid (2 points), hypoechoic (2 points), edge irregularity (2 points), macrocalcification (1 point), a total of 7 points, rated TR5, is a malignant nodule. Because of 1 point of macrocalcification, the ACR guidelines get the correct diagnosis. **d** This is a solid, very hypoechoic, irregularly shaped nodule. The pathological diagnosis is a malignant nodule. The ATA guideline is classified as highly suspicious and is a malignant nodule. The ACR guidelines are classified as solid (2 points), very hypoechoic (3 points), edge irregularity (2 points), a total of 7 points, rated TR5, is a malignant nodule. Because of 1 added point of very hypoechoic, the ACR guidelines get the correct diagnosis. **e** This is a solid, isoechoic, punctate calcification, irregular edge nodules. The pathological diagnosis is a malignant nodule. The ACR guideline is classified as TR5 and is a malignant nodule. The ATA guidelines cannot be classified. This is because the stratification of the ATA guidelines does not include the evaluation of hyperechoic nodules or isoechoic nodules with malignant signs (taller-than-wide, punctate, marginal irregularities, etc.)

performed when following the ATA guidelines, of which 348 were malignant and 234 were benign (35.2%). In the subgroup analysis, in which the TR5 of the ACR guidelines was compared with the highly suspicious group of the ATA guidelines and the TR4 of the ACR guidelines was compared with the intermediate suspicion group of the ATA guidelines, the differences were significant ($P < 0.05$) (Table 5). This is because the ATA guidelines have a higher weight given to marginal lobulation or infiltration, leading to more benign nodules being evaluated as highly

suspicious nodules, which in turn leads to more unnecessary FNA biopsies being performed. Moreover, the ACR guidelines increase the threshold for FNA biopsies in moderately suspicious nodules, from 1.0 cm (ATA guidelines) to 1.5 cm, which also reduces the number of unnecessary FNA biopsies performed. However, we also recognize that reducing the number of unnecessary FNA biopsies may reduce the detection of malignant nodules. In the 359 malignant nodules larger than 1 cm, 327 (91.1%) malignant nodules were detected according to the FNA biopsy standard of the ACR guidelines, and 348 (96.9%) malignant nodules were detected according to the FNA biopsy standard of the ATA guidelines. However, the difference between the two guidelines is not significant ($P = 0.180$).

Nodules <1 cm in diameter

Our study included nodules <1 cm in diameter, which is necessary to manage some patients with high-risk factors, such as childhood exposure history, familial thyroid cancer, lymphadenopathy, rapid nodule growth, and hoarseness. Although most current guidelines recommend active surveillance of nodules <1 cm exhibiting highly suspicious US features, we chose to perform biopsies to obtain a confirmed diagnosis. First, a confirmed diagnosis of benign nodules can reduce unnecessary active surveillance. Second, a confirmed diagnosis can help to alleviate the psychological burden on the patients. Finally, a confirmed diagnosis is required for some patients with high-risk factors even when the nodules are <1 cm in diameter. In addition, when a nodule <1 cm in diameter is diagnosed as a malignant nodule, we recommend that patients with high-risk factors undergo surgery, whereas those without high-risk factors were recommended to receive active surveillance. In this study, among 171 malignant nodules <1 cm, 163 nodules were actively surveilled, and the 8 nodules with high-risk factors were surgically excised. Thus, we also avoided overtreatment. For the subgroup of nodules <1 cm in diameter, compared with the ACR guidelines, the ATA guidelines still had a higher sensitivity and a larger NPV, while comparing with the ATA guidelines, the ACR guidelines had a higher specificity and a larger PPV; the diagnostic accuracies of the two guidelines were not significantly different. Previous studies have used the ACR guidelines to analyze nodules <1 cm [18, 19, 22], but Weiss's research did not compare the ACR and ATA guidelines. Though it compared the ACR and ATA guidelines, Middleton's study did not separately analyze nodules <1 cm in diameter. Only Ha EJ's study referred to the diagnostic performance of the sets of guidelines for nodules <1 cm in diameter, but this was not their main research aim. Our study analyzed 336 nodules <1 cm in diameter and found that the ATA guidelines still had a higher sensitivity and a larger NPV, while the ACR guidelines had a higher specificity and a larger PPV. However, the diagnostic accuracies of the two guidelines were not significantly different. This is slightly different from the results of the analysis of all 1000 nodules. We believe that the difference in accuracy for the two guidelines was not significant due to the small sample size.

Interobserver variability

Different from other studies on interobserver variability [28, 29], we take a completely different approach to interobserver variability. In our study, the three experts analyzed each nodule and then immediately discussed their findings to achieve a consensus before moving on to analyzing the next nodule. If the three experts disagreed, the opinion of the majority was followed. After the consensus was reached for a previous nodule, the agreement was used as a standard for subsequent analyses. With the help of this method, we have resolved the disagreements between the observers and the solutions to these disagreements based on expert discussions. For example, (1) it was unclear in distinguishing between the peripheral calcification and macrocalcification at the edge in the ACR guidelines. (2) When the echogenicity of nodule was hypoechoic but very close to the adjacent thyroid parenchyma, should this nodule be defined as isoechoic or hypoechoic? (3) In evaluating the margins, sometimes it is difficult to distinguish ill-defined margins from irregular margins. Experts achieved consensuses on how to solve all of these problems. However, the consensuses were only based on the understanding of the guidelines and the discussion of the three experts in our institution. It may not reflect the actual intentions of the guideline creators. Therefore, it is necessary to collect and explain these disagreements in future research.

Limits of this study

Our study had several limitations. First, the sample selection was biased. Before biopsy, all patients underwent US examination by an experienced sonographer. Only a nodule with highly suspicious US features was selected for biopsy. In addition, our institution is a tertiary referral center that treats patients with more severe disease. Thus, sample selection bias may have occurred, causing the higher malignancy rate of 53%, which is a rate much higher than those of unselected series of thyroid nodules. It also explains why 60.5 and 76.2% of the nodules were classified as high-risk according to the ACR and ATA guidelines, respectively, which were also much higher than expected. Second, the number of nodule samples <1 cm in diameter was small, and a larger sample size is needed in future

analyses. Third, for nodules <1 cm in diameter, a 12-month follow-up with decreased or stable size may not be sufficient to confirm the nodule as begin. Fourth, the clinical risk factors indicating the need for FNA biopsy are unknown and may need further research in the future. Fifth, since our study is based on cytology reports for a large number of nodules, both false negatives and false positives are possible.

Conclusion

This study suggested that the ACR and ATA guidelines had different strengths with regard to their diagnostic performance. The ATA guidelines are more sensitive than the ACR guidelines due to including the nodule margins as a critical grading indicator. The ACR guidelines are more accurate than the ATA guidelines because they include macrocalcification and very hypoechoic nodules. In terms of avoiding unnecessary FNA biopsies, the ACR guidelines significantly reduce the number due to an increased FNA threshold. The clinician can give optimal recommendations to patients with thyroid nodules based on the actual condition of the patient and the characteristics of the two guidelines.

Acknowledgements This study was funded by the Finance Department of Jilin Province (no. SCZSY201701), the Jilin Provincial Health and Family Planning Commission (no. 2016ZC032), and Jilin Province Science and Technology Department (no. 20170414042GH).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. H. Lim, S.S. Devesa, J.A. Sosa, D. Check, C.M. Kitahara, Trends in thyroid cancer incidence and mortality in the United States, 1974-2013. JAMA **317**(13), 1338–1348 (2017). https://doi.org/10.1001/jama.2017.2719
2. S. Vaccarella, S. Franceschi, F. Bray, C.P. Wild, M. Plummer, L. Dal Maso, Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. N. Engl. J. Med. **375**(7), 614–617 (2016). https://doi.org/10.1056/NEJMp1604412
3. S. Leboulleux, R.M. Tuttle, F. Pacini, M. Schlumberger, Papillary thyroid microcarcinoma: time to shift from surgery to active surveillance? Lancet Diabetes Endocrinol. **4**(11), 933–942 (2016). https://doi.org/10.1016/s2213-8587(16)30180-2
4. R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2018. CA Cancer J. Clin. **68**(1), 7–30 (2018). https://doi.org/10.3322/caac.21442
5. C. Durante, G. Grani, L. Lamartina, S. Filetti, S.J. Mandel, D.S. Cooper, The diagnosis and management of thyroid nodules: a review. JAMA **319**(9), 914–924 (2018). https://doi.org/10.1001/jama.2018.0898
6. E.K. Kim, C.S. Park, W.Y. Chung, K.K. Oh, D.I. Kim, J.T. Lee, H.S. Yoo, New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. AJR Am. J. Roentgenol. **178**(3), 687–691 (2002). https://doi.org/10.2214/ajr.178.3.1780687
7. H. Gharib, E. Papini, R. Paschke, D.S. Duick, R. Valcavi, L. Hegedus, P. Vitti, American Association of Clinical Endocrinologists, Associazione Medici Endocrinologi, and European Thyroid Association medical guidelines for clinical practice for the diagnosis and management of thyroid nodules. J. Endocrinol. Invest. **33**(5 Suppl), 1–50 (2010)
8. M.C. Frates, C.B. Benson, J.W. Charboneau, E.S. Cibas, O.H. Clark, B.G. Coleman, J.J. Cronan, P.M. Doubilet, D.B. Evans, J. R. Goellner, I.D. Hay, B.S. Hertzberg, C.M. Intenzo, R.B. Jeffrey, J.E. Langer, P.R. Larsen, S.J. Mandel, W.D. Middleton, C.C. Reading, S.I. Sherman, F.N. Tessler, Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. Radiology **237**(3), 794–800 (2005). https://doi.org/10.1148/radiol.2373050220
9. J.Y. Kwak, K.H. Han, J.H. Yoon, H.J. Moon, E.J. Son, S.H. Park, H.K. Jung, J.S. Choi, B.M. Kim, E.K. Kim, Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. Radiology **260**(3), 892–899 (2011). https://doi.org/10.1148/radiol.11110206
10. B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, F. Pacini, G.W. Randolph, A.M. Sawka, M. Schlumberger, K.G. Schuff, S.I. Sherman, J.A. Sosa, D.L. Steward, R.M. Tuttle, L. Wartofsky, 2015 American Thyroid Association Management Guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines Task Force on thyroid nodules and differentiated thyroid cancer. Thyroid **26**(1), 1–133 (2016). https://doi.org/10.1089/thy.2015.0020
11. T. Xu, J.Y. Gu, X.H. Ye, S.H. Xu, Y. Wu, X.Y. Shao, D.Z. Liu, W.P. Lu, F. Hua, B.M. Shi, J. Liang, L. Xu, W. Tang, C. Liu, X. H. Wu, Thyroid nodule sizes influence the diagnostic performance of TIRADS and ultrasound patterns of 2015 ATA guidelines: a multicenter retrospective study. Sci. Rep. **7**, 43183 (2017). https://doi.org/10.1038/srep43183
12. J.H. Yoon, K. Han, E.K. Kim, H.J. Moon, J.Y. Kwak, Diagnosis and management of small thyroid nodules: a comparative study with six guidelines for thyroid nodules. Radiology **283**(2), 560–569 (2017). https://doi.org/10.1148/radiol.2016160641
13. M. Boehnke, N. Patel, K. McKinney, T. Clark, Diagnostic performance of SRU and ATA thyroid nodule classification algorithms as tested with a 1 million virtual thyroid nodule model. Curr. Probl. Diagn. Radiol. **47**(1), 10–13 (2018). https://doi.org/10.1067/j.cpradiol.2017.04.005
14. F.N. Tessler, W.D. Middleton, E.G. Grant, J.K. Hoang, L.L. Berland, S.A. Teefey, J.J. Cronan, M.D. Beland, T.S. Desser, M. C. Frates, L.W. Hammers, U.M. Hamper, J.E. Langer, C.C. Reading, L.M. Scoutt, A.T. Stavros, ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-

RADS committee. J. Am. Coll. Radiol. **14**(5), 587–595 (2017). https://doi.org/10.1016/j.jacr.2017.01.046
15. Y. Zheng, S. Xu, H. Kang, W. Zhan, A. Single-Center Retrospective, Validation study of the American College of Radiology Thyroid Imaging Reporting and Data System. Ultrasound Q. **34**(2), 77–83 (2018). https://doi.org/10.1097/ruq.0000000000000350
16. J.K. Hoang, W.D. Middleton, A.E. Farjat, J.E. Langer, C.C. Reading, S.A. Teefey, N. Abinanti, F.J. Boschini, A.J. Bronner, N. Dahiya, B.S. Hertzberg, J.R. Newman, D. Scanga, R.C. Vogler, F.N. Tessler, Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. Radiology **287**(1), 185–193 (2018). https://doi.org/10.1148/radiol.2018172572
17. G. Grani, L. Lamartina, V. Ascoli, D. Bosco, M. Biffoni, L. Giacomelli, M. Maranghi, R. Falcone, V. Ramundo, V. Cantisani, S. Filetti, C. Durante, Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the "right" TIRADS. J. Clin. Endocrinol. Metab. **104**(1), 95–102 (2019). https://doi.org/10.1210/jc.2018-01674
18. E.J. Ha, D.G. Na, W.J. Moon, Y.H. Lee, N. Choi, Diagnostic performance of ultrasound-based risk-stratification systems for thyroid nodules: comparison of the 2015 American Thyroid Association Guidelines with the 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology and 2017 American Congress of Radiology Guidelines. Thyroid **28**(11), 1532–1537 (2018). https://doi.org/10.1089/thy.2018.0094
19. V.L. Weiss, R.F. Andreotti, K.A. Ely, Use of the thyroid imaging, reporting, and data system (TI-RADS) scoring system for the evaluation of subcentimeter thyroid nodules. Cancer Cytopathol. **126**(8), 518–524 (2018). https://doi.org/10.1002/cncy.22015
20. F.D. Koseoglu Atilla, B. Ozgen Saydam, Does the ACR TI-RADS scoring allow us to safely avoid unnecessary thyroid biopsy? single center analysis in a large cohort. Endocrine **61**(3), 398–402 (2018). https://doi.org/10.1007/s12020-018-1620-6
21. A. Lauria Pantano, E. Maddaloni, S.I. Briganti, G. Beretta Anguissola, E. Perrella, C. Taffon, A. Palermo, P. Pozzilli, S. Manfrini, A. Crescenzi, Differences between ATA, AACE/ACE/AME and ACR TI-RADS ultrasound classifications performance in identifying cytological high-risk thyroid nodules. Eur. J. Endocrinol. **178**(6), 595–603 (2018). https://doi.org/10.1530/eje-18-0083
22. W.D. Middleton, S.A. Teefey, C.C. Reading, J.E. Langer, M.D. Beland, M.M. Szabunio, T.S. Desser, Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association Guidelines. AJR Am. J. Roentgenol. **210**(5), 1148–1154 (2018). https://doi.org/10.2214/ajr.17.18822
23. E.J. Ha, D.G. Na, J.H. Baek, J.Y. Sung, J.H. Kim, S.Y. Kang, US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. Radiology **287**(3), 893–900 (2018). https://doi.org/10.1148/radiol.2018171074
24. A.S. Griffin, J. Mitsky, U. Rawal, A.J. Bronner, F.N. Tessler, J.K. Hoang, Improved quality of thyroid ultrasound reports after implementation of the ACR thyroid imaging reporting and data system nodule lexicon and risk stratification system. J. Am. Coll. Radiol. **15**(5), 743–748 (2018). https://doi.org/10.1016/j.jacr.2018.01.024
25. F.N. Tessler, W.D. Middleton, E.G. Grant, Thyroid imaging reporting and data system (TI-RADS): a user's guide. Radiology **287**(1), 29–36 (2018). https://doi.org/10.1148/radiol.2017171240
26. E.G. Grant, F.N. Tessler, J.K. Hoang, J.E. Langer, M.D. Beland, L.L. Berland, J.J. Cronan, T.S. Desser, M.C. Frates, U.M. Hamper, W.D. Middleton, C.C. Reading, L.M. Scoutt, A.T. Stavros, S. A. Teefey, Thyroid ultrasound reporting lexicon: white paper of the ACR thyroid imaging, reporting and data system (TIRADS) committee. J. Am. Coll. Radiol. **12**(12 Pt A), 1272–1279 (2015). https://doi.org/10.1016/j.jacr.2015.07.011
27. J.H. Yoon, H.S. Lee, E.K. Kim, H.J. Moon, J.Y. Kwak, Malignancy risk stratification of thyroid nodules: comparison between the thyroid imaging reporting and data system and the 2014 American Thyroid Association Management Guidelines. Radiology **278**(3), 917–924 (2016). https://doi.org/10.1148/radiol.2015150056
28. G. Grani, L. Lamartina, V. Cantisani, M. Maranghi, P. Lucia, C. Durante, Interobserver agreement of various thyroid imaging reporting and data systems. Endocr. Connect. **7**(1), 1–7 (2018). https://doi.org/10.1530/ec-17-0336
29. J. Koh, S.Y. Kim, H.S. Lee, E.K. Kim, J.Y. Kwak, H.J. Moon, J. H. Yoon, Diagnostic performances and interobserver agreement according to observer experience: a comparison study using three guidelines for management of thyroid nodules. Acta Radiol. (Stockholm, Sweden: 1987) **59**(8), 917–923 (2018). https://doi.org/10.1177/0284185117744001