Diagn Interv Radiol 2021; 27:511-518

HEAD AND NECK IMAGING

ORIGINAL ARTICLE

Diagnostic performance rates of the ACR-TIRADS and EU-TIRADS based on histopathological evidence

İlhan Hekimsoy
Egemen Öztürk
Yeşim Ertan
Mehmet Nurullah Orman
Gülgün Kavukçu
Ahmet Gökhan Özgen
Murat Özdemir
Süha Süreyya Özbek

PURPOSE
In this study, we aimed to assess the effectiveness of malignancy stratification algorithms of the American College of Radiology (ACR) and European Thyroid Association (ETA) in the delineation of thyroid nodules using a database of nodules that were unequivocally diagnosed by means of histopathological examination and meticulously matched with the imaged nodules.

METHODS
A total of 165 patients having 251 thyroid nodules with histopathologically proven definitive diagnoses during a 5-year period were included in this study. All patients had preoperatively undergone ultrasonography (US) examination, and US characteristics of the thyroid nodules were retrospectively analyzed and assigned in compliance with the thyroid imaging reporting and data system categories recommended by the ACR (ACR-TIRADS) and ETA (EU-TIRADS). The diagnostic effectiveness in the delineation of thyroid nodules and unnecessary fine-needle aspiration (FNAB) rates were evaluated.

RESULTS
Overall, 189 nodules (75.30%) were diagnosed as benign, while 62 nodules (24.70%) were reported to be malignant based on histopathological assessment. Sensitivity and specificity rates were 71% and 75% for ACR-TIRADS and 73% and 80% for EU-TIRADS. The area under the curve values were 0.78 and 0.80 for ACR-TIRADS and EU-TIRADS, respectively. The unnecessary FNAB rates were 61% for ACR-TIRADS and 64% for EU-TIRADS as per the recommended criteria of each algorithm.

CONCLUSION
The diagnostic performance of both malignancy stratification systems was signified to be moderate and sufficient in a cohort of nodules with definite histopathological diagnosis. In light of our results, we demonstrated the strengths and weaknesses of the ACR- and EU-TIRADS for physicians who should be familiar with them for optimal management of thyroid nodules.

The number of detected thyroid nodules has been increasing in recent years with the widespread application of ultrasonography (US). The reported incidence varies from 20% to 68% in patients undergoing high-frequency US examination (1, 2). In case of a newly detected nodule, the primary concern is to discriminate benign ones, which constitute almost 90%, from malignant ones that require additional invasive procedures (3, 4). Fine-needle aspiration biopsy (FNAB) is the primary diagnostic tool due to its high sensitivity and specificity in distinguishing malignancy, yet it has several shortcomings, including inconclusive results and potential overdiagnosis (5).

The particular nodular US features suggestive of malignancy are well known and thus US is employed for the indication of FNAB (6, 7). Nevertheless, none of those characteristics individually predicts the malignancy risk sufficiently (8, 9). Hence, various risk-stratification systems that consider a set of nodular US features have been established to predict the malignancy risk, mitigate superfluous FNABs, and enhance interobserver concordance. Among them, the Thyroid Imaging Reporting and Data System of the American College of Radiology (ACR-TIRADS) was set up in 2017 to classify all detected thyroid nodules according to its issued lexicon (10–12). Similarly, the European Thyroid Association TIRADS

From the Departments of Radiology (İ.H. ✉ *ihekimsoy@hotmail.com, ihekimsoy@gmail.com*, E.Ö., G.K., S.S.Ö.), Pathology (Y.E.), Biostatistics and Medical Informatics (M.N.O.), Internal Medicine (A.G.Ö.), and General Surgery (M.Ö.), Ege University Faculty of Medicine, İzmir, Turkey.

Received 2 October 2020; revision requested 12 November 2020; last revision received 20 January 2021; accepted 15 February 2021.

Published online 8 April 2021.

DOI 10.5152/dir.2021.20813

You may cite this article as: Hekimsoy İ, Öztürk E, Ertan Y, et al. Diagnostic performance rates of the ACR-TIRADS and EU-TIRADS based on histopathological evidence. Diagn Interv Radiol 2021; 27:511-518

(EU-TIRADS) was developed to improve the sensitivity together with high negative predictive value (NPV) in characterization with a more straightforward scoring method (12, 13). Several studies have compared the effectiveness of these two risk-stratification systems (14–18), yet their performance in different thyroid nodules with various histopathologic results and subtypes still needs to be investigated.

The objectives of our study were to appraise the diagnostic effectiveness of the ACR- and EU-TIRADS classification systems in nodular characterization and to analyze the rates of inappropriate FNABs according to the proposed criteria based on unequivocal histopathological results.

Methods

Study population

This retrospective study was approved by the institutional review board of our institution (decision number: 18-5/17), and any requirement of informed consent was waived.

The list of 2447 patients who had undergone 7660 detailed US examinations of the thyroid gland during a 5-year period was obtained to compose a study population with histopathologically evaluated thyroid nodules. Their names and institutional patient identity numbers were compared with the same parameters of 4399 patients who had undergone thyroidectomy and had a histopathological diagnosis for nodule(s) in the same time interval (Fig. 1). In total, 220 patients (with 601 consecutive US examinations) were present in both lists, thus having at least one nodule with histopathological diagnosis.

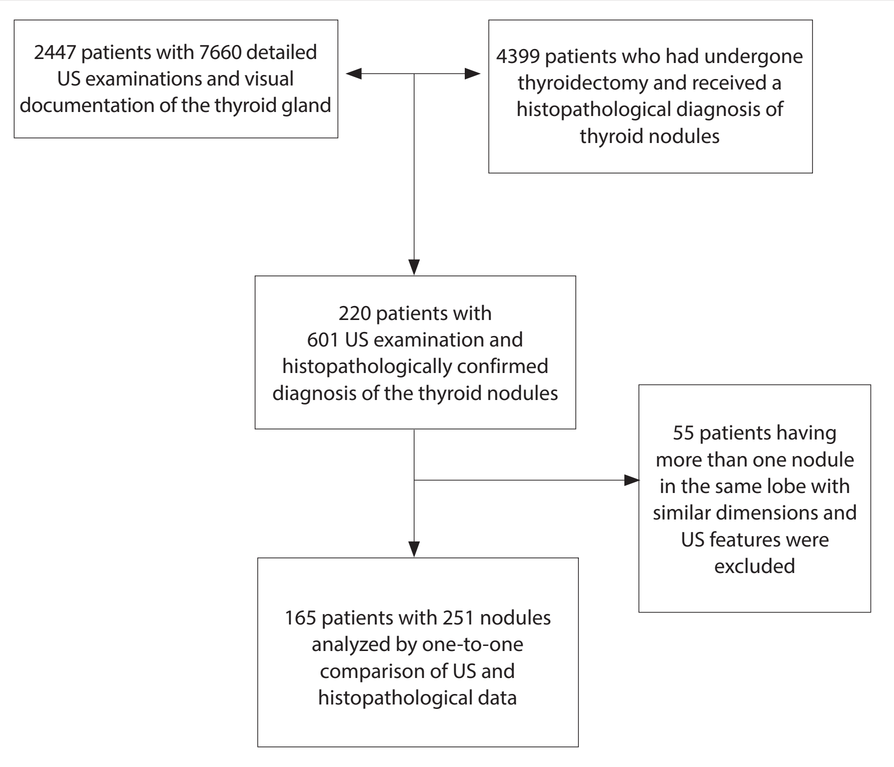
2447 patients with 7660 detailed US examinations and visual documentation of the thyroid gland
4399 patients who had undergone thyroidectomy and received a histopathological diagnosis of thyroid nodules
220 patients with 601 US examination and histopathologically confirmed diagnosis of the thyroid nodules
55 patients having more than one nodule in the same lobe with similar dimensions and US features were excluded
165 patients with 251 nodules analyzed by one-to-one comparison of US and histopathological data

Figure 1. The flowchart of the study population.

Main points

- Using a one-to-one matching approach in determining the definite histopathological diagnoses of thyroid nodules, our study demonstrated moderate, similar, and acceptable rates of diagnostic performance of the ACR-TIRADS and EU-TIRADS risk classification systems with AUC values of 0.78 and 0.80, respectively.
- The ICC values were calculated as 0.94 (95% CI, 0.92–0.95; $p < 0.001$) for the ACR-TIRADS and 0.90 (95% CI, 0.88–0.92; $p < 0.001$) for the EU-TIRADS.
- Overall, 16% of the nodules in the ACR-TIRADS and 15% of the nodules in the EU-TIRADS, designated as benign or assigned to routine surveillance according to the original recommendations, were diagnosed as malignant in final histopathological examinations.
- Both risk stratification systems can be useful in radiological diagnosis and clinical management of thyroid nodules, keeping in mind their advantages and drawbacks.

The one-to-one matching of nodular histopathological diagnoses with the corresponding nodules was performed meticulously according to the following process. Originally defined US features and images of the nodules in the latest preoperative US examination were compared with the histopathological characteristics of the thyroid nodule(s) in pathology reports, which had been separately produced for each thyroideal lobe. The distinctive features of nodules such as laterality, dimensions, and calcification were used to assure that an attributed malignant histopathological diagnosis belongs to a specific nodule defined in the preoperative US examination. In case of a nodule with malignant histopathological diagnosis in one lobe having multiple nodules with similar dimensions and US features, the nodule and its lobe were excluded from the study to avoid any doubt about matching. Besides, any thyroideal lobe histopathologically reported to harbor a microcarcinoma was also excluded unless it had been detected in the preoperative US exam. On the contrary, all the nodules in the thyroid lobes which were histopathologically reported to harbor no malignant focus were accepted as benign and included in the study group if their US features had been described in detail and their sonographic image was available.

Thus, 251 nodules (in 165 patients) were assured of having a definite histopathological diagnosis. Of them, 189 benign nodules (75.30%) were located in thyroideal lobes with the diagnosis of follicular nodular disease, without any demonstrated focus of malignancy. The remaining 62 nodules (24.70%) were diagnosed as malignant. All except one (with medullary carcinoma) were reported to be papillary thyroid carcinoma (PTC). As for the 61 nodules with PTC, 26 nodules (42.62%) were follicular, 22 (36.07%) were classical type, 10 (16.39%) were oncocytic, one (1.64%) was solid, and one (1.64%) was tall cell variant. Histopathologic information regarding the microscopic variant was not available for one PTC nodule (1.64%).

Ultrasound examination

Three senior radiologists with more than 20 years of US experience separately performed all examinations using the Acuson S1000$^{\text{TM}}$ or S2000$^{\text{TM}}$ US devices (Siemens

Medical Solutions) equipped with 9L4 (4–9 MHz) in addition to 14L5 (5–14 MHz) or 18L6 (5.5–18 MHz) linear transducers. US examinations were conducted with the patient in supine position with the neck gently extended. The echogenicity of the thyroid parenchyma and the presence of any intraglandular nodules were initially evaluated. Dimensions, echogenicity, border regularity, internal composition, as well as the presence of punctate hyperechogenic foci (colloid crystals, micro-/macrocalcifications or undefined) and extra-thyroidal extension of the nodules were assessed according to accepted definitions and reported in detail (11, 13). At least one descriptive image of each significant nodule was obtained. These still images were digitally archived, along with detailed examination reports.

Assessment of stored images

Each of the 251 nodules was randomly numbered. The most recent preoperative US report and image(s) of each nodule were anonymized and grouped in a separate digital folder containing the designated number as the folder name. Thus, using 251 digital folders created in this way, three researchers independently and retrospectively classified each nodule according to the ACR- and EU-TIRADS systems. One of the researchers was an experienced radiologist with 26 years of US experience. The second one was a young radiologist who had recently completed residency, and the last one was a third-year radiology resident. Subsequently, all researchers gathered to compare their judgments and to reach a consensus on the classification of each discordantly stratified nodule. The diagnostic performance rates of the ACR- and EU-TIRADS systems were investigated based on the histopathological diagnoses of nodules. Finally, nodules requiring FNAB according to the proposed criteria of each system were also identified to determine the number of unnecessary biopsies, which in turn would yield benign biopsy results given the histopathological outcomes. Furthermore, we performed an additional trial of comparison in this manner. To enhance the comparability of both systems in terms of nodular length, the cutoff values of mildly, moderately, and highly suspicious categories of both TIRADS systems for FNAB were equalized. For this purpose, we reanalyzed the ratio of unnecessary biopsy for the EU-TIRADS system when the size thresholds of EU-TIRADS were hypothetically changed to the values equaling the size criteria of the ACR-TIRADS system (i.e., FNAB, if the nodular length is ≥2.5 cm for EU-TIRADS 3, ≥1.5 cm for EU-TIRADS 4, and ≥1.0 cm for EU-TIRADS 5 lesions).

Statistical analysis

As for the test of normality, Kolmogorov–Smirnov analysis was performed. Throughout the manuscript, all variables without normal distribution are reported as median (Q1–Q3, 25th–75th percentile values). On the other hand, all normally distributed variables are depicted as mean values (±standard deviation, SD). The categorical variables are reported as number (percentage). After appropriate descriptive analyses, the independent samples t-test or Mann–Whitney U test was performed to compare the groups depending on whether the data had normal distribution or not, respectively. The Spearman's rank correlation coefficient was calculated to detect whether any possible relationship exists between the US and histopathological features of malignant nodules regarding the maximal diameters. In accordance with statistical requirements, Fisher Freeman Halton, Fisher exact, or Pearson chi-square test was used to evaluate any association of malignant nature of nodules with gender of the patients and the US characteristics of nodules. Stepwise backward logistic regression analysis was performed to calculate odds ratios and to determine the significance of the relationship between different sonographic features and malignant histopathologic results of the studied nodules. Interobserver agreement among three researchers was evaluated, calculating the intraclass correlation coefficient (ICC) for each malignancy stratification system. The diagnostic power of both classification systems in detecting malignant nodules was explored based on their sensitivity, specificity, positive predictive value (PPV), NPV, and positive and negative likelihood ratios. The receiver operator characteristic (ROC) curve analyses were performed, and the areas under the curves (AUCs) were calculated to compare the diagnostic accuracy of both algorithms. The head-to-head comparison of the AUCs was achieved using the same method as DeLong et al. (19). All statistical analyses were performed using IBM-SPSS 25.0 for Windows software package (IBM Corp.). A p value was considered to be statistically significant when <0.05.

Results

The study was based on 251 histopathologically confirmed thyroid nodules, detected in 165 patients with a mean age of 49.64±13.50 years (95% CI, 47.56–51.71 years; range, 17–78 years) at the time of their most recent preoperative US scan. The detailed demographic features of the study population, along with the US characteristics of nodules, are presented in Table 1. Of the patients, 34 (20.60%) were male (mean age, 53.29±13.49 years; range, 19–76 years) with 46 nodules (18.33%) and 131 (79.40%) were female (mean age, 48.69±13.39 years; range, 17–78 years) with 205 nodules (81.67%). The numbers of male and female patients with at least one malignant nodule were calculated as 12 (35.29% of males) and 47 (35.88% of females), respectively. Mean age ($p = 0.076$) and rate of malignant nodules ($p = 0.950$) were not significantly different among male and female patients. However, patients with at least one malignant nodule were considerably younger than those without ($p = 0.021$) (Table 1).

The distribution of nodules in the study according to their longest diameters was as follows: <5 mm (1 benign, 1 malignant), 5–10 mm (49 benign, 16 malignant), 11–15 mm (43 benign, 15 malignant), 16–20 mm (30 benign, 4 malignant), and >20 mm (66 benign, 26 malignant). The longest sonographic diameters of the benign nodules were between 4.00 and 57.00 mm with a median value of 16.00 mm (Q1–Q3, 10.00–27.50 mm), while the malignant nodules had the longest diameters ranging from 3.00 to 85.00 mm with a median length of 15.00 mm (10.00–29.25 mm). No remarkable difference was found between the longest sonographic diameters of the nodules with malignant or benign histopathology (Table 1). Meanwhile, the longest histopathologic diameters of 62 malignant nodules obtained from the pathology reports varied between 3.00 and 45.00 mm (median, 14.00 mm; Q1–Q3, 8.00–25.25 mm), which were not statistically different from sonographically obtained diameters and strongly correlated with them (r=0.92, $p < 0.001$).

Hypoechogenicity (either mildly or markedly), irregular borders, "taller-than-wide" shape, and internal microcalcifications were significantly more frequent in malignant nodules, while no similar association was observed with the presence of the remain-

Table 1. Demographic and ultrasound features of the study population and nodules

	Histopathological result			
Parameter	Benign	Malignant	Total	*p*
Number of patients	106 (64)	59 (36)	165	
Gender[a]				0.950
Male	22 (21)	12 (20)	34	
Female	84 (79)	47 (80)	131	
Patient age (years), mean±SD	51.43±12.53	46.41±14.64		0.021
Number of nodules	189 (75)	62 (25)	251	
Nodule size (mm), median (Q1–Q3)	16.0 (10.0–27.5)	15.0 (10.0–29.25)		0.700
Composition, n (%)[b]				0.145
Cystic or almost completely cystic	2 (1.1)	0 (0)	2	
Spongiform	1 (0.5)	0 (0)	1	
Mixed cystic and solid	36 (19)	5 (8)	41	
Solid or almost completely solid	150 (79.4)	57 (92)	207	
Echogenicity, n (%)[b]				<0.001*
Anechoic	2 (1)	0 (0)	2	
Hyper or isoechoic	156 (83)	22 (35)	178	
Hypoechoic	31 (16)	37 (60)	68	
Very hypoechoic	0 (0)	3 (5)	3	
Shape, n (%)[c]				<0.001
Wider-than-tall	187 (99)	45 (73)	232	
Taller-than-wide	2 (1)	17 (27)	19	
Margins, n (%)[c]				<0.001
Smooth or ill-defined	187 (99)	33 (53)	220	
Lobulated or irregular	2 (1)	29 (47)	31	
Calcifications, n (%)				
Comet-tail artifacts[a]	37 (20)	6 (10)	43	0.073
Macrocalcification[c]	11 (6)	8 (13)	19	0.093
Microcalcification[c]	3 (1.6)	11 (18)	14	<0.001

Data are presented as n (%) unless otherwise indicated.
SD, standard deviation; Q1–Q3, 25th–75th percentiles.
[a]Pearson chi-square test; [b]Fisher Freeman Halton; [c]Fisher exact test.
*For the sake of statistical robustness, two anechoic (cystic) nodules were excluded from this chi-squared test, while three markedly hypoechoic nodules were included in the hypoechoic group and analyzed together in a single group.

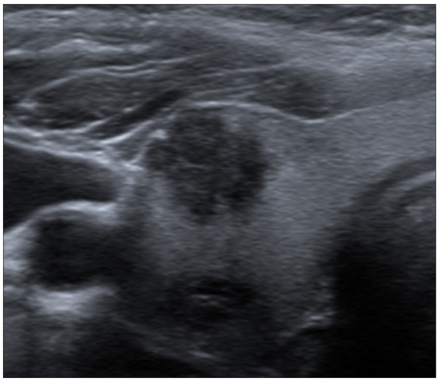

Figure 2. Transverse thyroideal US scan of a 50-year-old woman demonstrates a markedly hypoechoic solid nodule with border irregularity. The nodule with the histopathologic diagnosis of papillary carcinoma (classical type) was blindly and unanimously graded as "highly suspicious" (TIRADS 5) with ACR- and EU-TIRADS.

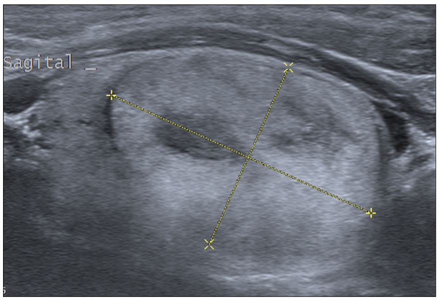

Figure 3. Sagittal sonographic image of the right thyroideal lobe in a 40-year-old male. The cursors mark a 35 mm long solid and predominantly isoechoic nodule with relatively small hypoechoic areas. This nodule was one of those on which the grading of three raters was discordant regarding echogenicity and eventually graded as a "mildly suspicious" nodule (TIRADS 3) according to both systems in the consensus meeting. The final histopathologic diagnosis for the lobe was follicular nodular disease without any sign of a malignant lesion.

ing US features on a nodule-based analysis, nor with the gender of patients (Fig. 2; Table 1). However, further analysis with stepwise backward logistic regression demonstrated that only hypoechogenicity and border irregularity of the nodules were independently associated with a significantly increased risk of malignancy (Supplemental Table 1).

A perfect interobserver agreement was noted among three researchers in classifying 251 nodules according to the ACR- and EU-TIRADS. The ICC values for both classification systems were calculated to be 0.94 (95% CI, 0.92–0.95; $p < 0.001$) and 0.90 (95% CI, 0.88-0.92; $p < 0.001$), respectively. The determined classes of both systems by three researchers were the same in 207 nodules (82.47%). At least one reviewer was discordant in 35 nodules (13.94%) among the ACR-TIRADS classes, 29 nodules (11.55%) among the EU-TIRADS classes, and 20 nodules (7.97%) among classes of both systems (Fig. 3). In this context, 44 nodules (17.53%) with discordant classes were reassessed to establish the consensus category. The most frequent causes of discordance are highlighted in Table 2. The distribution of the final scores, according to the consensus meeting, is yielded in Table 3.

In the present study, only TIRADS 5 categories of both systems were shown to have calculated malignancy ratios compatible with the originally proposed risk rates (11, 13). The escalation of malignancy rates was observed as the TIRADS categories increased in both systems, and their performance rates were not significantly different. Based on final histopathological results, the diagnostic performance rates of both systems in two different settings, namely TIRADS 4-5 and TIRADS 5 categories, are summarized in Table 4.

In a recent histopathological reevaluation, 5 of 26 nodules (19.23%) with the diagnosis of follicular variant PTC were classified as "noninvasive follicular thyroid neoplasm with papillary-like nuclear features" (NIFTP), which was recommended to be accepted as an indolent tumor in 2017 (20). The recalculated performance rates of the two systems are demonstrated in Table 5 with these nodules assumed to have a benign histopathologic diagnosis.

Table 2. The most frequent and classifiable causes of discordance between reviewers

Cause of controversies	ACR-TIRADS	EU-TIRADS	Total
Deciding for "border irregularity" in nodules with minimal undulated contours (especially in large ones)	2	3	5
Defining nodules with few tiny cystic spaces as "solid" or "mixed"	16	4	20
Interpreting predominantly isoechoic nodules with internal heterogeneity as "isoechoic" or "hypoechoic" when they partly have hypoechoic areas	12	14	26
Defining nodules with homogeneous and minimally hypoechoic internal appearance as "isoechoic" or "hypoechoic"	2	2	4
Determining punctate hyperechogenic foci as microcalcification or colloid crystals	0	3	3

Data are shown as numbers.

Table 3. The distribution of nodules' final scores, according to the consensus meeting

Classification system	Malignant (n=62)	Benign (n=189)	Total (n=251)
ACR-TIRADS			
1	0 (0)	3 (100)	3
2	4 (9)	42 (91)	46
3	14 (13)	97 (87)	111
4	19 (31)	43 (69)	62
5	25 (86)	4 (14)	29
EU-TIRADS			
1	0 (0)	0 (0)	0
2	0 (0)	3 (100)	3
3	17 (10)	148 (90)	165
4	13 (30)	31 (70)	44
5	32 (82)	7 (18)	39

Data are presented as numbers (percentages).
ACR-TIRADS, thyroid imaging reporting and data system of the American College of Radiology; EU-TIRADS, thyroid imaging reporting and data system of the European Thyroid Association.

The ROC analysis revealed AUCs of 0.784 (95% CI, 0.728–0.833) for the ACR-TIRADS and 0.803 (95% CI, 0.748–0.850) for the EU-TIRADS when NIFTP considered as malignant. However, accepting NIFTP as a benign pathology yielded AUCs of 0.810 (95% CI, 0.756–0.857) and 0.830 (95% CI, 0.777–0.784) for the ACR- and EU-TIRADS, respectively. On further analysis, a head-to-head comparison of these AUCs demonstrated no significant difference in diagnostic performances of the risk-stratification methods when NIFTP was regarded as malignant (p = 0.213; 95% CI, -0.011 to 0.049; SE 0.015) or benign (p = 0.171; 95% CI, -0.008 to 0.047; SE 0.014).

Finally, to test performance rates of the biopsy suggestions of both classification systems, potentially unnecessary intervention rates were determined by calculating the ratios of histopathologically confirmed benign nodules that would have been referred to FNAB according to the originally proposed recommended criteria of both systems. These ratios were 61% for the ACR-TIRADS and 64% for the EU-TIRADS. On further analysis, after the size criteria of the ACR-TIRADS for biopsy were applied in the EU-TIRADS group, we found that the unnecessary FNAB ratio of the EU-TIRADS decreased from 64% to 61%, precisely the same as that of the ACR-TIRADS. On the other hand, 16% of the nodules in ACR-TIRADS and 15% of the nodules in EU-TIRADS, designated as benign or assigned to routine surveillance according to the original recommendations, were diagnosed as malignant in final histopathological examinations.

Discussion

Contrary to its high accuracy in detection, US is not equally successful in characteriz-

Table 4. Diagnostic performances of ACR-TIRADS and EU-TIRADS based on different cutoff categories

	Malignant (n=62)	Benign (n=189)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	PLR	NLR
ACR-TIRADS 1–3	18	142	71 (58–82)	75 (68–81)	48 (38–59)	89 (83–93)	2.9 (2.1–3.8)	0.4 (0.3–0.6)
ACR-TIRADS 4–5	44	47						
EU-TIRADS 1–3	17	151	73 (60–83)	80 (73–85)	54 (43–65)	90 (84–94)	3.6 (2.6–4.9)	0.3 (0.2–0.5)
EU-TIRADS 4–5	45	38						
ACR-TIRADS 1–4	37	185	40 (28–54)	98 (95–99)	86 (68–96)	83 (78–88)	19 (7–53)	0.6 (0.5–0.8)
ACR-TIRADS 5	25	4						
EU-TIRADS 1–4	30	182	52 (39–65)	96 (93–99)	82 (66–92)	86 (80–90)	14 (6–30)	0.5 (0.4–0.7)
EU-TIRADS 5	32	7						

Data are presented as numbers or percentages. The numbers in parentheses are 95% confidence intervals.
ACR-TIRADS, thyroid imaging reporting and data system of the American College of Radiology; EU-TIRADS, thyroid imaging reporting and data system of the European Thyroid Association; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

Table 5. Diagnostic performances of ACR-TIRADS and EU-TIRADS based on different cutoff categories when NIFTP is accepted as a benign pathology

	Malignant (n=57)	Benign (n=194)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	PLR	NLR
ACR-TIRADS 1–3	13	147	77 (64–87)	76 (69–82)	48 (38–59)	92 (87–96)	3.2 (2.4–4.2)	0.3 (0.2–0.5)
ACR-TIRADS 4–5	44	47						
EU-TIRADS 1–3	13	155	77 (64–87)	80 (74–85)	53 (42–64)	92 (87–96)	3.8 (2.8–5.3)	0.3 (0.2–0.5)
EU-TIRADS 4–5	44	39						
ACR-TIRADS 1–4	32	190	44 (31–58)	98 (95–99)	86 (68–96)	86 (80–90)	21 (8–59)	0.6 (0.5–0.7)
ACR-TIRADS 5	25	4						
EU-TIRADS 1–4	25	187	56 (42–69)	96 (93–99)	82 (66–92)	88 (83–92)	16 (7–33)	0.5 (0.3–0.6)
EU-TIRADS 5	32	7						

Data are presented as numbers or percentages. The numbers in parentheses are 95% confidence intervals.
ACR-TIRADS, thyroid imaging reporting and data system of the American College of Radiology; EU-TIRADS, thyroid imaging reporting and data system of the European Thyroid Association; NIFTP, noninvasive follicular thyroid neoplasm with papillary-like nuclear features; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

ing thyroid nodules (21). Thus, several study groups, including the ACR and ETA, constructed algorithms to ameliorate the diagnostic performance using a combination of US features. The experience and validation related to them in daily practice are still growing. Unlike many of the previous work validating the ACR- and EU-TIRADS systems based mainly on cytopathologic results (14, 15, 18, 22–25), our study was conducted solely using unequivocal histopathological diagnoses with microscopic variant analysis of PTC, thereby highlighting the impact of follicular variant PTC and NIFTP on the effectiveness of these risk-stratification systems. Given the limitations of FNAB, we believe that this approach provides a different and more solid basis for scientific discussion (5, 26–29).

The current study could only entail the most recently developed malignancy stratification systems (i.e., ACR- and EU-TIRADS) considering the significant number of disparate US-based risk-stratification algorithms. Moreover, the EU-TIRADS system was selected for its relative ease of use, whereas the ACR-TIRADS was preferred for its unique approach based on total point-scoring. These choices also enabled us to compare the diagnostic effectiveness of two discrete US-based approaches. The ACR-TIRADS weighs the US characteristics according to their malignancy risks by scoring them from 0 to 3, while the EU-TIRADS uses a more straightforward approach by denoting some US features a high risk of malignancy, i.e., pattern-based approach. Apart from this significant difference, these two algorithms have some variations in their terminology and definitions. For instance, the presence of macrocalcification and rim calcification is not considered in the EU-TIRADS, while they have certain additive points for malignancy risk of a nodule in the ACR-TIRADS. The ACR-TIRADS accepts comet-tail artifact as a feature suggesting benignity only if it has a posterior tail longer than 1 mm, unlike the EU-TIRADS system, which assumes echogenic foci in a nodular cystic component with all sizes of comet-tail artifacts as benign colloid crystals. In contrast to the EU-TIRADS, mixed nodular composition yields a lower score in the ACR-TIRADS, which may lead to underestimating the TIRADS score of a malignant nodule with a cystic component. These issues partly accounted for the interrater discordance in grading nodules in the presented study (Table 2). Finally, the nodular size thresholds for FNAB recommendation differ, most remarkably for "mildly suspicious" (low-risk) nodules in both systems as 2.0 vs. 2.5 cm for the EU-TIRADS and ACR-TIRADS, respectively.

In the current study, the majority of patients that had undergone thyroidectomy were female. In line with the report by Shen et al. (16), no significant difference was found in our study between males and females in terms of nodules' malignancy ratio, while patients with malignant nodules were significantly younger. Similar to the results of Frates et al. (30), no remarkable difference was demonstrated between the long diameters of malignant and benign nodules. Although irregular borders, "taller-than-wide" shape, and microcalcifications were significantly more prevalent in malignant nodules in accordance with previous academic work, stepwise backward logistic regression analysis demonstrated that irregular margins and hypoechogenicity were the only independent predictors of malignancy in our series (21).

For both risk-stratification systems compared in our study, interobserver reliability analysis indicated excellent consistency among the three reviewers having different experience levels. Conversely, for nodule classification, Grani et al. (14) found a better interobserver agreement with the EU-TIRADS than with the ACR-TIRADS, according to the ratings of two appraisers with the same background who evaluated the data sets before and after specific training. In light of our results displaying the reasons for discordance between observers, we infer that more precise definitions, particularly in determining irregular borders, characterizing nodules with few tiny cystic spaces, and assessment of nodules having heterogeneous echotexture would reinforce the interreader agreement in both systems.

The calculated malignancy risks of the ACR-TIRADS 1 and EU-TIRADS 2 categories were found to be 0%, which aligned with the study by Brito et al. (21), indicating cystic or spongiform nodules as benign. In the remaining categories except TIRADS 5, we calculated higher malignancy ratios than those determined for both models (11, 13). We assume that the enrollment of only histopathologically confirmed nodules with one-to-one matching might have led to this discrepancy, given the potential problems of matching nodular US findings with their histopathological results in multinodular thyroid lobes, and particularly the well-known risk of false-negative diagnosis in FNAB (27, 28).

The diagnostic performances of these two classification systems were similar according to our analyses accepting TIRADS

4 and 5 categories as malignant. Moreover, assuming only the TIRADS 5 category as malignant yielded further increment in specificity, PPV, and positive likelihood ratio of both algorithms while lowering sensitivity and NPV. In our opinion, the relatively high number (42.62%) of follicular variant PTC, most of which were depicted as an isoechoic solid nodule with smooth margins and thus were categorized as ACR- and EU-TIRADS 3, gave rise to lower sensitivity and NPV in our cohort. On the other hand, the classification of NIFTP as a benign pathology improved sensitivity, NPV, and positive likelihood ratio slightly regardless of the cutoff categories, while not changing specificity and PPV notably. None of the nodules reclassified as NIFTP had highly suspicious US features. Three of them were assigned to TIRADS 3 category according to both risk-stratification methods. Only one with a mild hypoechoic solid component was designated as EU-TIRADS 4, whereas owing to its mixed cystic and solid composition, its ACR-TIRADS category was 3. Similar to our results, Chaigneau et al. (31) depicted that accepting NIFTP as a benign pathology increased the NPV of the French TIRADS model for the categories of TIRADS 3 and 4A.

For both classification systems, unnecessary FNAB and malignancy rates among nodules assigned to routine surveillance were quite similar in our study. Given that the definition of unnecessary FNAB rate differs in various studies, some of which defined it as the number of benign nodules in the total enrolled nodules or nodules measuring more than 1 cm (18, 32, 33), it is difficult to compare our results with the previous studies. However, we believe that it is more appropriate to describe this rate as the number of benign nodules in the FNAB-recommended nodules, as some authors did (34, 35). In our study, we found lower unnecessary FNAB rates than the study by Huh et al. (35), who reported this value as 63.8% and 73.9% for the ACR- and EU-TIRADS, respectively. This discrepancy between outcomes could be ascribed to the difference of the gold standard method between studies, as Huh et al. (35) determined benignity via FNAB in most nodules. On the other hand, aligned with the study of Huh et al. (35), indicating the reduction of the unnecessary FNAB ratios of the EU-TIRADS from 73.6% to 70.6% after applying the size threshold of the ACR-TIRADS, we also detected a noteworthy decline in the unnecessary FNAB ratios of the EU-TIRADS using the same hypothetical scenario. In contradiction to our results, Grani et al. (15) reported better outcomes with the ACR-TIRADS recommendations than those of EU-TIRADS in determining deferrable FNABs with the calculated malignancy rates of 2.2% and 3.2%, respectively. Conversely, Xu et al. (18) detected much higher malignancy ratios in nodules, of which FNAB was deferrable according to the ACR- and EU-TIRADS criteria: 33.1% and 37.7%, respectively. These discrepancies probably stemmed from differences in the prevalence of malignant nodules among the studies, which was reported to be 7.2% by Grani et al. (15) and 40% by Xu et al. (18). Supporting this argument, malignancy rates of deferrable biopsies were 16% for ACR-TIRADS and 15% for EU-TIRADS recommendations in our series, with a malignancy rate of 24.7%. Besides, these two papers were based on cytopathology results and dimensional stability in the follow-up studies to determine whether a nodule is benign or not. However, these two criteria are not always reliable to exclude malignancy, so we believe that this approach might be prone to type II error (26–29, 36–38).

Our study has several limitations. First, the retrospective study design led to the interpretation of static images instead of real-time ones, which may have impeded the TIRADS classification of nodules. However, meticulous optimization of imaging settings, in addition to the detailed evaluation and description of nodular US features available in examination reports, should have minimized this shortcoming. Second, due to the one-to-one matching process of the nodules with histopathological diagnoses performed in the study, we could enroll a relatively small number of nodules, and the ratio of malignant nodules (24.7%) was relatively high compared to that in the general population. However, our malignancy ratio was similar to other histopathology-based studies, such as the ones by Borlea et al. (17) and Trimboli et al. (39). Furthermore, we believe that unequivocal histopathological confirmation enabled us to assess the real diagnostic performances, considering the possibility of false-negative results in cytopathologic analysis and histopathological matching problems in multinodular thyroid lobes. Third, given the fact that almost all of the malignant nodules in our study were PTC, we could not evaluate the effectiveness of classification systems in other pathological types of thyroid cancer. Nevertheless, analyzing microscopic variants of PTC allowed us to demonstrate their effect on diagnostic performances, strengthening the scientific value of our study. In this context, further prospective studies exploring the diagnostic accuracy of malignancy-stratification methods in different forms of thyroid cancer appear to be needed.

In conclusion, both ACR- and EU-TIRADS showed acceptable and similar performances in predicting thyroid nodule malignancy risks with perfect interobserver reliability regardless of their experience levels. However, considering the relatively high unnecessary FNAB rates in our study, refinement of biopsy criteria seems to be needed for both algorithms. Therefore, radiologists and clinicians should be aware of the advantages and drawbacks of these two classification systems while incorporating them into the management of thyroid nodules.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Guth S, Theune U, Aberle J, et al. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest 2009; 39:699–706. [Crossref]
2. Fisher SB, Perrier ND. The incidental thyroid nodule. CA Cancer J Clin 2018; 68:97–105. [Crossref]
3. Davies L, Welch HG. Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg 2014; 140:317–322. [Crossref]
4. Burman KD, Wartofsky L. Clinical practice. Thyroid nodules. N Engl J Med 2015; 373:2347–2356. [Crossref]
5. Grani G, Calvanese A, Carbotta G, et al. Intrinsic factors affecting adequacy of thyroid nodule fine-needle aspiration cytology. Clin Endocrinol (Oxf) 2013; 78:141–144. [Crossref]
6. Kim E-K, Park CS, Chung WY, et al. New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. AJR Am J Roentgenol 2002; 178:687–691. [Crossref]
7. Bonavita JA, Mayo J, Babb J, et al. Pattern recognition of benign nodules at ultrasound of the thyroid: which nodules can be left alone? AJR Am J Roentgenol 2009; 193:207–213. [Crossref]
8. Campanella P, Ianni F, Rota CA, et al. Quantification of cancer risk of each clinical and ultrasonographic suspicious feature of thyroid nodules: a systematic review and meta-analysis. Eur J Endocrinol 2014; 170:R203–211. [Crossref]
9. Remonti LR, Kramer CK, Leitão CB, et al. Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. Thyroid 2015; 25:538–550. [Crossref]

10. Grant EG, Tessler FN, Hoang JK, et al. Thyroid Ultrasound Reporting Lexicon: White Paper of the ACR Thyroid Imaging, Reporting and Data System (TIRADS) Committee. J Am Coll Radiol 2015; 12(12 Pt A):1272–1279. [Crossref]
11. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017; 14:587–595. [Crossref]
12. Tappouni RR, Itri JN, McQueen TS, et al. ACR TI-RADS: pitfalls, solutions, and future directions. Radiographics 2019; 39:2040–2052. [Crossref]
13. Russ G, Bonnema SJ, Erdogan MF, et al. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: The EU-TIRADS. Eur Thyroid J 2017; 6:225–237. [Crossref]
14. Grani G, Lamartina L, Cantisani V, et al. Interobserver agreement of various thyroid imaging reporting and data systems. Endocr Connect 2018; 7:1–7. [Crossref]
15. Grani G, Lamartina L, Ascoli V, et al. Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the "right" TIRADS. J Clin Endocrinol Metab 2019; 104:95–102. [Crossref]
16. Shen Y, Liu M, He J, et al. Comparison of different risk-stratification systems for the diagnosis of benign and malignant thyroid nodules. Front Oncol 2019; 9:378. [Crossref]
17. Borlea A, Borcan F, Sporea I, et al. TI-RADS diagnostic performance: which algorithm is superior and how elastography and 4D vascularity improve the malignancy risk assessment. Diagnostics (Basel) 2020; 10:180. [Crossref]
18. Xu T, Wu Y, Wu R-X, et al. Validation and comparison of three newly-released Thyroid Imaging Reporting and Data Systems for cancer risk determination. Endocrine 2019; 64:299–307. [Crossref]
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–845. [Crossref]
20. Nikiforov YE, Ghossein RA, Kakudo K, et al. Non-invasive follicular thyroid neoplasm with papillary-like nuclear features. In: Lloyd RV, Osamura RY, Klöppel G, Rosai J, ed. WHO Classification of Tumours of Endocrine Organs. 4th ed. Lyon: IARC Press, 2017:78–80.
21. Brito JP, Gionfriddo MR, Al Nofal A, et al. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. J Clin Endocrinol Metab 2014; 99:1253–1263. [Crossref]
22. Maino F, Forleo R, Martinelli M, et al. Prospective validation of ATA and ETA sonographic pattern risk of thyroid nodules selected for FNAC. J Clin Endocrinol Metab 2018; 103:2362–2368. [Crossref]
23. Clark TJ, McKinney K, Jensen A, et al. Risk threshold algorithm for thyroid nodule management demonstrates increased specificity and diagnostic accuracy as compared with American College of Radiology Thyroid Imaging, Reporting and Data System; Society of Radiologists in Ultrasound; and American Thyroid Association Management Guidelines. Ultrasound Q 2019; 35:224–227. [Crossref]
24. Modi L, Sun W, Shafizadeh N, et al. Does a higher American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) score forecast an increased risk of malignancy? A correlation study of ACR TI-RADS with FNA cytology in the evaluation of thyroid nodules. Cancer Cytopathol 2020; 128:470–481. [Crossref]
25. Di Fermo F, Sforza N, Rosmarin M, et al. Comparison of different systems of ultrasound (US) risk stratification for malignancy in elderly patients with thyroid nodules. Real world experience. Endocrine 2020; 69:331–338. [Crossref]
26. Yang GC, Liebeskind D, Messina AV. Ultrasound-guided fine-needle aspiration of the thyroid assessed by Ultrafast Papanicolaou stain: data from 1135 biopsies with a two- to six-year follow-up. Thyroid 2001; 11:581–589. [Crossref]
27. Tee YY, Lowe AJ, Brand CA, et al. Fine-needle aspiration may miss a third of all malignancy in palpable thyroid nodules: a comprehensive literature review. Ann Surg 2007; 246:714–720. [Crossref]
28. Pinchot SN, Al-Wagih H, Schaefer S, et al. Accuracy of fine-needle aspiration biopsy for predicting neoplasm or carcinoma in thyroid nodules 4 cm or larger. Arch Surg 2009; 144:649–655. [Crossref]
29. Bojunga J, Herrmann E, Meyer G, et al. Real-time elastography for the differentiation of benign and malignant thyroid nodules: a meta-analysis. Thyroid 2010; 20:1145–1150. [Crossref]
30. Frates MC, Benson CB, Doubilet PM, et al. Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. J Clin Endocrinol Metab 2006; 91:3411–3417. [Crossref]
31. Chaigneau E, Russ G, Royer B, et al. TIRADS score is of limited clinical value for risk stratification of indeterminate cytological results. Eur J Endocrinol 2018; 179:13–20. [Crossref]
32. Yoon SJ, Na DG, Gwon HY, et al. Similarities and differences between thyroid imaging reporting and data systems. AJR Am J Roentgenol 2019; 213:W76–84. [Crossref]
33. Wu X-L, Du J-R, Wang H, et al. 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. Endocrine 2019; 65:121–131. [Crossref]
34. Ruan J-L, Yang H-Y, Liu R-B, et al. Fine needle aspiration biopsy indications for thyroid nodules: compare a point-based risk stratification system with a pattern-based risk stratification system. Eur Radiol 2019; 29:4871–4878. [Crossref]
35. Huh S, Lee HS, Yoon J, et al. Diagnostic performances and unnecessary US-FNA rates of various TIRADS after application of equal size thresholds. Sci Rep 2020; 10:10632. [Crossref]
36. Clark TJT, Pokharel S, Meier J, et al. Thyroid nodule doubling time is not a reliable indicator of benign or malignant nature. Ultrasound Q 2016; 32:132–135. [Crossref]
37. Singh Ospina N, Maraka S, Espinosa DeYcaza A, et al. Diagnostic accuracy of thyroid nodule growth to predict malignancy in thyroid nodules with benign cytology: systematic review and meta-analysis. Clin Endocrinol (Oxf) 2016; 85:122–131. [Crossref]
38. Cordes M, Götz TI, Horstrup K, et al. Growth rates of malignant and benign thyroid nodules in an ultrasound follow-up study: a retrospective cohort study. BMC Cancer 2019; 19:1139. [Crossref]
39. Trimboli P, Ngu R, Royer B, et al. A multicentre validation study for the EU-TIRADS using histological diagnosis as a gold standard. Clin Endocrinol (Oxf) 2019; 91:340–347. [Crossref]

Table S1. Relationship between different US features and malignant histopathologic results based on stepwise backward logistic regression analysis

		Stepwise backward LR	
Sonographic features	Malignancy rate (%)	OR (95% CI)	*p*
Long diameter	N/A	1.05 (1.02–1.07)	0.001
Composition*			
Mixed	5/41 (12)		
Solid	57/206 (28)		
Echogenicity*		5.54 (2.44–12.56)	<0.001
Isoechoic	22/176 (13)		
Hypoechoic	40/71 (56)		
Shape			
Wider-than-tall	45/228 (20)		
Taller-than-wide	17/19 (90)		
Borders		64.68 (13.25–315.68)	<0.001
Regular	33/216 (15)		
Irregular	29/31 (94)		
Punctuate echogenic foci			
Absent	43/189 (23)		
Present	19/58 (33)		
Macrocalcification			
Absent	54/228 (24)		
Present	8/19 (42)		
Microcalcification			
Absent	51/233 (22)		
Present	11/14 (79)		
Colloid crystal		2.63 (0.82–8.42)	0.10
Absent	56/206 (27)		
Present	6/41 (15)		

US, ultrasonography; LR, logistic regression; OR, odds ratio; 95% CI, confidence interval; N/A, non-applicable.
Model *p* value ≤0.001 (Omnibus test). Cox & Snell $R^2 = 0.351$; Nagelkerke $R^2 = 0.519$.
*Due to their extensively limited numbers, two cystic, one spongiform, and one hyperechoic solid nodules were excluded from the logistic regression analyses, while all hypoechoic nodules, including mild and marked hypoechoic ones, were analyzed together in a single group.