Original Article | Thyroid

eISSN 2005-8330
https://doi.org/10.3348/kjr.2019.0215
Korean J Radiol 2020;21(1):108-116

Korean Journal of Radiology
KJR

Concordance of Three International Guidelines for Thyroid Nodules Classified by Ultrasonography and Diagnostic Performance of Biopsy Criteria

Younghee Yim, MD[1], Dong Gyu Na, MD, PhD[2], Eun Ju Ha, MD, PhD[3], Jung Hwan Baek, MD, PhD[4], Jin Yong Sung, MD[5], Ji-hoon Kim, MD, PhD[6], Won-Jin Moon, MD, PhD[7]

[1]Department of Radiology, College of Medicine, Kangwon National University, Kangwon National University Hospital, Chuncheon, Korea; [2]Department of Radiology, GangNeung Asan Hospital, University of Ulsan College of Medicine, Gangneung, Korea; [3]Department of Radiology, Ajou University, School of Medicine, Suwon, Korea; [4]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea; [5]Department of Radiology, Thyroid Center, Daerim St. Mary's Hospital, Seoul, Korea; [6]Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Korea; [7]Department of Radiology, Konkuk University Medical Center, Konkuk University School of Medicine, Seoul, Korea

Objective: To investigate the concordance of three international guidelines: the Korean Thyroid Association/Korean Society of Thyroid Radiology, American Thyroid Association, and American College of Radiology for thyroid nodules classified by ultrasonography (US) and the diagnostic performance of simulated size criteria for malignant biopsies.

Materials and Methods: A total of 2586 thyroid nodules (≥ 1 cm) were collected from two multicenter study datasets. The classifications of the thyroid nodules were based on three different guidelines according to US categories for malignancy risk, and the concordance rate between the different guidelines was calculated for the classified nodules. In addition, the diagnostic performance of criteria related to four different simulated biopsy sizes was evaluated.

Results: The concordance rate of nodules classified as high- or intermediate-suspicion was high (84.1–100%), but low-suspicion or mildly-suspicious nodules exhibited relatively low concordance (63.8–83.8%) between the three guidelines. The differences in sensitivity, specificity, and accuracy between the guidelines were 0.7–19.8%, 0–40.9%, and 0.1–30.5%, respectively, when the original biopsy criteria were applied. The differences decreased to 0–5.9%, 0–10.9%, and 0.1–8.2%, respectively, when simulated, similar biopsy size criteria were applied. The unnecessary biopsy rate calculated with the original criteria (0–33.8%), decreased with the simulated biopsy size criteria (0–8.7%).

Conclusion: We found a high concordance between the three guidelines for high- or intermediate-suspicion nodules, and the diagnostic performance of the biopsy criteria was approximately equivalent for each simulated size criterion. The difference in diagnostic performance between the three guidelines is mostly influenced by the various size thresholds for biopsies.

Keywords: *Thyroid nodules; Thyroid cancer; Thyroid imaging reporting and data system; Ultrasonography; Size threshold; Guidelines*

INTRODUCTION

Many international guidelines have recently been proposed to provide management recommendations for various ultrasonography (US)-based thyroid nodule risk-stratification systems (1-6). The American Thyroid Association (ATA) guidelines (2) recommend fine-needle aspiration (FNA) for high- or intermediate- (≥ 1 cm), low- (≥ 1.5 cm), and very-low- suspicion (≥ 2 cm) nodules. The Korean Thyroid Association/Korean Society of Thyroid

Received March 31, 2019; accepted after revision October 14, 2019.
Corresponding author: Dong Gyu Na, MD, PhD, Department of Radiology, GangNeung Asan Hospital, University of Ulsan College of Medicine, 38 Bangdong-gil, Sacheon-myeon, Gangneung 25440, Korea.
• Tel: (8233) 610-4310 • Fax: (8233) 610-3490 • E-mail: nndgna@gmail.com

Korean Journal of Radiology
KJR

Radiology (KTA/KSThR) guidelines (4) recommend FNA for similar size thresholds, i.e., high-, intermediate-, and low-suspicion nodules. The American College of Radiology (ACR) guidelines (5) recommend FNA for highly-suspicious nodules ≥ 1 cm and suggest higher FNA size thresholds for moderately-suspicious (≥ 1.5 cm) and mildly-suspicious nodules (≥ 2.5 cm). It does not recommend FNA for non-suspicious or benign category nodules. Although various international guidelines require different US categorization structures, size thresholds for biopsy, and diagnostic performance of FNA criteria for detection of thyroid malignancy (7-11), few studies (12) have investigated factors influencing the diagnostic performance between the guidelines.

Therefore, the purpose of this study was to evaluate the concordance of the three international guidelines for US classified nodules and the diagnostic performance of biopsy size criteria for malignancy using simulated biopsy size thresholds.

MATERIALS AND METHODS

The Institutional Review Boards of participating centers approved this multicenter study. Informed consent was waived for the retrospective study.

Study Population

Patient data were retrospectively collected from two previously published multicenter-study datasets (13, 14). Dataset 1 included 2000 thyroid nodules (≥ 1 cm) diagnosed with a pathology examination, from 1802 consecutive patients (1387 women and 415 men, mean age 51.2 years, age range 13–79 years). Dataset 2 included 586 thyroid nodules (≥ 1 cm) diagnosed with a pathology examination, from 499 consecutive patients (396 women and 103 men, mean age 50.4 years, age range 10–81 years). The final diagnosis for thyroid nodules in both datasets was determined by surgery, and the final diagnosis for benign nodules was determined by a pathology examination of the surgical results, FNA, or core-needle biopsy (CNB) (15, 16). Nodules with non-diagnostic or indeterminate biopsy results without surgical confirmation were excluded. We intentionally included two datasets with different patient populations because the distribution of US classified nodules and the diagnostic performance of biopsy criteria might be affected by the proportion and histological type of malignant tumors in a certain population.

US Examination and Image Analysis

All US examinations were performed using a 10–16 MHz linear probe and various real-time US systems. We analyzed databases from two previously published studies (9, 10), comparing thyroid guidelines based on two different cohort datasets (13, 14). Dataset 1 US images were retrospectively assessed by one of three radiologists, each with 12–19 years of experience in thyroid US (9). Dataset 2 US features were prospectively assessed by one of five radiologists, specializing in thyroid imaging, each with 8–20 years of experience in thyroid US (10). Several US features, including configuration of solid components in partially cystic nodules and extrusive soft-tissue components in rim calcification, were retrospectively assessed in the two datasets (9, 10).

US nodule features, such as internal content, echogenicity, margin, calcification, or shape (orientation), were assessed as described in previous studies (9, 10). In partially cystic nodules, the configuration of solid areas was categorized as concentric or eccentric. In nodules with rim calcification, the presence or absence of an extrusive soft-tissue component was subdivided for analysis using ATA guidelines. The US feature, "extrathyroidal extension (ETE)," included in the ATA and ACR guidelines, was not analyzed due to the absence of standardized US criteria for the diagnosis of ETE. Unclassified nodules according to the ATA guidelines were categorized as intermediate-suspicion nodules, based on previous studies (9, 10, 17). We simulated four different size criteria (criteria 1–4), to compare the diagnostic performance of biopsy criteria, according to similar size thresholds.

Statistical Analysis

Nodules were retrospectively classified according to categories defined by the KTA/KSThR, ATA, and ACR guidelines. The demographic data from the two datasets were compared by using an unpaired *t* test for numerical data (age and nodule size) and a chi-square test or Fisher's exact test for categorical data (sex, size distribution, proportion of malignant tumors, and histological type). The proportion of classified nodules and histological types of malignant tumors in the two datasets was determined with the chi-square test or Fisher's exact test.

The correlation of classified nodules between the different guidelines was calculated using the Spearman's correlation test. The concordance rate of classified nodules was defined as the concordance rate between the different guidelines in the same risk level categories. McNemar's test was applied to determine the concordance rate of diagnostic performance

 https://doi.org/10.3348/kjr.2019.0215

for simulated and similar biopsy size criteria, assessed by comparing the sensitivity, specificity, accuracy, and unnecessary biopsy rates between the different guidelines. The unnecessary biopsy rate for the diagnosis of thyroid malignancy was defined as the number of benign nodules among the nodules requiring FNA.

All statistical analyses were performed using SPSS for Windows version 23.0 (IBM Corp., Armonk, NY, USA). The definition for significant differences was $p < 0.05$.

RESULTS

Demographic Data and Comparison of Two Datasets

Table 1 presents comparative results between the two datasets. The proportion of malignant tumors was significantly different between dataset 1 (22.7%, 454/2000) and dataset 2 (17.2%, 101/586) ($p = 0.005$). The proportion of papillary thyroid cancer (PTC) was significantly lower in dataset 1 than in dataset 2 (85.5%, 388/454 vs. 95%, 96/101, respectively, $p = 0.009$), and the proportion of follicular thyroid cancer (FTC) was higher in dataset 1 than that in dataset 2 (10.6%, 48/454 vs. 5.0%, 5/101, respectively), but without statistical significance ($p = 0.082$). The size of malignant tumors was not significantly different between the two datasets ($p = 0.611$); for example, malignant nodules of 1.0 to 1.4 cm were found in 52.9% (240/454) of dataset 1, and 51.5% (52/101) of dataset 2.

Table 1. Comparison of Demographic Data between Two Datasets of Thyroid Nodules (≥ 1 cm)

Data Characteristics	Dataset 1 (n = 2000)	Dataset 2 (n = 586)	*P*
Sex (female)	1387 (77.0)	457 (78.0)	0.610
Mean age, years (mean ± SD)	51.2 ± 12.2	50.3 ± 12.5	0.215
Nodule size, mm (mean ± SD)	20.0 ± 11.4	19.3 ± 11.2	0.123
Size distribution of all nodules			0.505
1.0–1.4 cm	835 (41.8)	246 (42.0)	
1.5–1.9 cm	450 (22.5)	141 (24.1)	
2.0–2.4 cm	226 (11.3)	72 (12.3)	
≥ 2.5 cm	489 (24.5)	127 (21.7)	
Malignant tumors	454 (22.7)	101 (17.2)	0.005
Papillary thyroid cancer	388 (85.5)	96 (95.0)	0.009
Follicular thyroid cancer	48 (10.6)	5 (5.0)	0.082
Size distribution of malignant tumors			0.611
1.0–1.4 cm	240 (52.9)	52 (51.5)	
1.5–1.9 cm	72 (15.9)	21 (20.8)	
2.0–2.4 cm	34 (7.5)	8 (7.9)	
≥ 2.5 cm	108 (23.8)	20 (19.8)	

Numbers in parentheses are percentages. SD = standard deviation

Table 2 shows the distribution of classified nodules within each dataset. Although the proportion of classified nodules was significantly different between the two datasets ($p < 0.001$), the proportion of low-suspicion or mildly-suspicious nodules was the highest (32.3–56.0%), followed by intermediate- or moderately-suspicious nodules (24.7–30.7%), in both datasets, with all three guidelines. The proportion of malignant tumors classified by US categories based on each of the three guidelines was significantly different between the two datasets (Table 3). The proportion of low- or very-low-suspicion US pattern malignant tumors based on the KTA/KSThR and ATA guidelines, was significantly higher in dataset 1 than that in dataset 2 (19.4%, 88/454 vs. 6.9%, 7/101, respectively, $p = 0.003$). The proportion of not- or mildly-suspicious US pattern tumors based on the ACR guidelines was significantly higher in dataset 1 than that in dataset 2 (16.7%, 76/454 vs. 6.9%, 7/101, respectively, $p = 0.013$) (Table 3). The proportion of FTC in dataset 1 among the malignant tumors with a low-suspicion pattern was significantly higher (23.9%, 21/88) than the proportion with an intermediate- or high-suspicion pattern (7.4%, 27/366), ($p < 0.001$), based on the KTA/KSThR guidelines.

Table 2. Comparison of Classified Thyroid Nodules (≥ 1 cm) by US Categories between Two Datasets in Each Guideline

US Pattern	Dataset 1 (n = 2000)	Dataset 2 (n = 586)	*P*
KTA/KSThR			< 0.001
Benign	53 (2.7)	41 (7.0)	
Low suspicion	1120 (56.0)	293 (50.0)	
Intermediate suspicion	533 (26.7)	157 (26.8)	
High suspicion	294 (14.7)	95 (16.2)	
ATA			< 0.001
Benign	0 (0.0)	0 (0.0)	
Very low suspicion	264 (13.2)	136 (23.2)	
Low suspicion	909 (45.5)	198 (33.8)	
Intermediate suspicion	506 (25.3)*	145 (24.7)*	
High suspicion	321 (16.1)	107 (18.3)	
ACR			< 0.001
Benign	38 (1.9)	23 (3.9)	
Not suspicious	273 (13.7)	93 (15.9)	
Mildly suspicious	768 (38.4)	189 (32.3)	
Moderately suspicious	614 (30.7)	180 (30.7)	
Highly suspicious	307 (15.4)	101 (17.2)	

Numbers in parentheses are percentages. *Unclassified nodules were categorized as intermediate suspicion nodules. ACR = American College of Radiology, ATA = American Thyroid Association, KTA/KSThR = Korean Thyroid Association/Korean Society of Thyroid Radiology, US = ultrasonography

https://doi.org/10.3348/kjr.2019.0215

Korean Journal of Radiology
KJR

Concordance of Thyroid Nodule Classified Categories between the KTA/KSThR, ATA, and ACR Guidelines

Classified nodules were highly correlated between the KTA/KSThR and ATA guidelines in datasets 1 and 2 (r = 0.937, r = 0.931, respectively), the KTA/KSThR and ACR guidelines in datasets 1, and 2 (r = 0.889, r = 0.902, respectively), and the ATA and ACR guidelines in datasets 1, and 2 (r = 0.912, r = 0.933, respectively), (all, $p < 0.001$).

Tables 4 and 5 list the concordance results between the KTA/KSThR, ATA, and ACR guidelines for the different thyroid nodule classification categories. Concordance rates between the KTA/KSThR and ATA, and between the KTA/KSThR and ACR guidelines for high- or highly-suspicious, and intermediate- or moderately-suspicious categories, were greater than 90% in both datasets. However, concordance rates between the KTA/KSThR and ATA guidelines for the low- and mildly-suspicious categories were relatively low in both datasets 1 and 2 (80.2% vs. 67.6%, respectively), and (68.3% vs. 63.8%, respectively) .

The concordance rate between the ATA and ACR guidelines for high- or highly-suspicious categories was 90.3% in dataset 1 and 84.1% in dataset 2. The intermediate- or moderately-suspicious category concordance rates in datasets 1 and 2 were 96.2% vs. 92.4%, respectively, and the low or mildly-suspicious category rates in datasets 1 and 2 were 79.5% vs. 83.8%, respectively.

Table 3. Comparison of Classified Malignant Tumors by US Categories between Two Datasets in Each Guideline

US Pattern	Dataset 1 (n = 454)	Dataset 2 (n = 101)	*P*
KTA/KSThR			0.004
Low suspicion	88 (19.4)	7 (6.9)	
Intermediate suspicion	133 (29.3)	27 (26.7)	
High suspicion	233 (51.3)	67 (66.3)	
ATA			0.001
Very low suspicion	13 (2.9)	4 (4.0)	
Low suspicion	75 (16.5)	3 (3.0)	
Intermediate suspicion	120 (26.4)	22 (21.8)	
High suspicion	246 (54.2)	72 (71.3)	
ACR			0.008
Not suspicious	12 (2.6)	4 (4.0)	
Mildly suspicious	64 (14.1)	3 (3.0)	
Moderately suspicious	141 (31.1)	28 (27.7)	
Highly suspicious	237 (52.2)	66 (65.3)	

Numbers in parentheses are percentages.

Diagnostic Performance of Biopsy Criteria according to Simulated Size Thresholds

Tables 6 and 7 present the diagnostic performance data of biopsy criteria for malignancy, based on the application of four different size criteria, in the two datasets. When the diagnostic performance using the original biopsy size criteria was compared between the three guidelines, the sensitivity between the KTA/KSThR and ATA guidelines was similar in both datasets, (dataset 1, 94.5% vs. 93.8%; dataset 2, 100%, vs. 99.0%, respectively); however, the sensitivity of the ACR guidelines in both datasets was significantly lower than that of the KTA/KSThR or ATA guidelines, (dataset 1, 74.7%; dataset 2, 80.2%; all, $p < 0.001$). When biopsy size criteria 1 or 2 were applied, there was no significant difference in sensitivity in both datasets between the three guidelines ($p \geq 0.250$). However, when

Table 4. Concordance of Classified Categories of Thyroid Nodules between KTA/KSThR, ATA, and ACR Guidelines (Dataset 1)

Guideline and Category	ATA					ACR				
	Benign	Very Low Suspicion	Low Suspicion	Intermediate Suspicion	High Suspicion	Benign	Not Suspicious	Mildly Suspicious	Moderately Suspicious	Highly Suspicious
KTA/KSThR										
Benign	0	42 (79.2)	11 (20.8)	0	0	36 (67.9)	13 (24.5)	3 (5.7)	1 (1.9)	0
Low suspicion	0	222 (19.8)	898 (80.2)	0	0	0	260 (23.2)	765 (68.3)	95 (8.5)	0
Intermediate suspicion	0	0	0	506 (94.9)	27 (5.1)	2 (0.4)	0	0	503 (94.4)	28 (5.3)
High suspicion	0	0	0	0	294 (100)	0	0	0	15 (5.1)	279 (94.9)
ATA										
Benign						0	0	0	0	0
Very low suspicion						36 (13.6)	179 (67.8)	45 (17.0)	4 (1.5)	0
Low suspicion						0	94 (10.3)	723 (79.5)	92 (10.1)	0
Intermediate suspicion						2	0	0	487 (96.2)	17 (3.4)
High suspicion						0	0	0	31 (9.7)	290 (90.3)

Data are number of nodules (% of nodules in each category of KTA/KSThR, ATA, or ACR guidelines) in dataset 1 (n = 2000).

https://doi.org/10.3348/kjr.2019.0215

Table 5. Concordance of Classified Categories of Thyroid Nodules between KTA/KSThR, ATA, and ACR Guidelines (Dataset 2)

Guideline and Category	ATA					ACR				
	Benign	Very Low Suspicion	Low Suspicion	Intermediate Suspicion	High Suspicion	Benign	Not Suspicious	Mildly Suspicious	Moderately Suspicious	Highly Suspicious
KTA/KSThR										
Benign	0	41 (100)	0	0	0	23 (56.1)	16 (39.0)	2 (4.9)	0	0
Low suspicion	0	95 (32.4)	198 (67.6)	0	0	0	77 (26.3)	187 (63.8)	29 (9.9)	0
Intermediate suspicion	0	0	0	145 (92.4)	12 (7.6)	0	0/0	0/0	143 (91.1)	14 (8.9)
High suspicion	0	0	0	0	95 (100)	0	0/0	0/0	8 (8.4)	87 (91.6)
ATA										
Benign						0	0	0	0	0
Very low suspicion						23 (16.9)	89 (65.4)	23 (16.9)	1 (0.7)	0
Low suspicion						0	4 (2.0)	166 (83.8)	28 (14.1)	0
Intermediate suspicion						0	0	0	134 (92.4)	11 (7.6)
High suspicion						0	0	0	17 (15.9)	90 (84.1)

Data are number of nodules (% of nodules in each category of KTA/KSThR, ATA, or ACR guidelines) in dataset 2 (n = 586).

Table 6. Simulated Biopsy Criteria for Biopsy Size Thresholds in KTA/KSThR, ATA, and ACR Guidelines

Simulated Biopsy Criteria	Simulated Size Thresholds for Biopsy according to US Categories of Thyroid Nodules
Criteria 1	
KTA/KSThR*	≥ 1 cm for high or intermediate suspicion nodules, ≥ 1.5 cm for low suspicion nodules
ATA*	≥ 1 cm for high or intermediate suspicion nodules, ≥ 1.5 cm for low suspicion nodules, ≥ 2 cm for very low suspicion nodules
ACR	≥ 1 cm for highly or moderately suspicious nodules, ≥ 1.5 cm for mildly suspicious nodules
Criteria 2	
KTA/KSThR	≥ 1 cm for high or intermediate suspicion nodules, ≥ 2 cm for low suspicion nodules
ATA	≥ 1 cm for high or intermediate suspicion nodules, ≥ 2 cm for low and very low suspicion nodules
ACR	≥ 1 cm for highly or moderately suspicion nodules, ≥ 2 cm for mildly suspicious nodules
Criteria 3	
KTA/KSThR	≥ 1 cm for high suspicion nodules, ≥ 1.5 cm for intermediate suspicion nodules, ≥ 2 cm for low suspicion nodules
ATA	≥ 1 cm for high suspicion nodules, ≥ 1.5 cm for intermediate suspicion nodules, ≥ 2 cm for low and very low suspicion nodules
ACR	≥ 1 cm for highly suspicious nodules, ≥ 1.5 cm for moderately suspicious nodules, ≥ 2 cm for mildly suspicious nodules
Criteria 4	
KTA/KSThR	≥ 1 cm for high suspicion nodules, ≥ 1.5 cm for intermediate suspicion nodules, ≥ 2.5 cm for low suspicion nodules
ATA	≥ 1 cm for high suspicion nodules, ≥ 1.5 cm for intermediate suspicion nodules, ≥ 2.5 cm for low and very low suspicion nodules
ACR*	≥ 1 cm for highly suspicious nodules, ≥ 1.5 cm for moderately suspicious nodules, ≥ 2.5 cm for mildly suspicious nodules

*Original biopsy size criteria of each guideline.

biopsy size criteria 3 or 4 were applied, the ATA guidelines provided a slightly higher sensitivity than that of the KTA/KSThR and ACR guidelines ($p = 0.004$, $p \geq 0.019$, respectively) in dataset 1, and a higher sensitivity than that of the ACR guidelines ($p = 0.031$) in dataset 2. The difference in sensitivity between the KTA/KSThR and ATA guidelines was 0.7–1.0% with the original biopsy size criteria, and 0–2.0% with the simulated, similar biopsy size criteria. Although the difference in sensitivity between the ACR and KTA/KSThR, or ATA guidelines was 18.8–19.8% with the original biopsy size criteria, the difference in sensitivity was decreased to 0–5.9% when the simulated, similar biopsy size criteria were applied.

The estimated specificity with the original biopsy size criteria was similar between the KTA/KSThR and ATA guidelines (dataset 1, 26.4% vs. 27.9%; dataset 2, 28.0%, vs. 28.0%, respectively); however, the specificity of the ACR guidelines was significantly higher than that of the KTA/KSThR or ATA guidelines in both datasets (dataset 1, 37.7%; dataset 2, 37.1%; all, $p < 0.001$). Specificity was significantly different between the three guidelines in dataset 1 when the four different criteria were applied ($p \leq 0.008$). In dataset 2, the biopsy size criteria 1 or 2 provided significantly different specificity between the KTA/KSThR or ATA and ACR guidelines ($p \leq 0.007$), while there was

https://doi.org/10.3348/kjr.2019.0215

Korean Journal of Radiology
KJR

Table 7. Diagnostic Performance of Simulated Biopsy Criteria in KTA/KSThR, ATA, and ACR Guidelines

Simulated Biopsy Criteria	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	Unnecessary FNA Rate (%)
Criteria 1						
KTA/KSThR*	94.5/100	26.4/28.0	27.4/22.4	94.2/100	41.9/40.4	56.9/59.6
ATA*	93.8/99.0	27.9/28.0	27.7/22.3	93.9/99.3	42.9/40.3	55.7/59.6
ACR	93.4/96.0	37.3/37.1	30.4/24.1	95.1/97.8	50.1/47.3	48.5/52.0
Criteria 2						
KTA/KSThR	90.5/97.0	43.3/43.9	31.9/26.5	94.0/98.6	54.1/53.1	43.8/46.4
ATA	90.5/97.0	41.7/40.4	31.3/25.3	93.8/98.5	52.8/50.2	45.1/49.3
ACR	90.3/94.1	48.4/48.9	34.0/27.7	94.5/97.5	58.0/56.7	39.9/42.3
Criteria 3						
KTA/KSThR	76.9/85.1	55.2/57.9	33.5/29.7	89.0/94.9	60.1/62.6	34.7/34.8
ATA	78.9/86.1	53.2/53.4	33.1/27.8	89.5/94.9	59.0/59.0	36.2/38.6
ACR	75.6/80.2	61.4/63.9	36.5/31.6	89.5/93.9	64.7/66.7	29.8/29.9
Criteria 4						
KTA/KSThR	75.6/85.1	63.7/66.4	37.9/34.5	89.9/95.5	66.4/69.6	28.1/27.8
ATA	77.5/86.1	62.2/62.7	37.6/32.5	90.4/95.6	65.7/66.7	29.2/30.9
ACR*	74.7/80.2	67.3/68.9	40.2/34.9	90.1/94.4	69.0/70.8	25.3/25.8

Data are diagnostic values of dataset 1 (n = 2000)/diagnostic values of dataset 2 (n = 586). *Original biopsy size criteria of each guideline. FNA = fine-needle aspiration, NPV = negative predictive value, PPV = positive predictive value

no significant difference between the KTA/KSThR and ATA guidelines ($p \geq 0.184$). When the biopsy size criteria 3 or 4 were applied, there was a significant difference in specificity between the three guidelines, except for the specificity between the KTA/KSThR and ACR guidelines, when the biopsy size criterion 4 was applied ($p = 0.193$). The difference in specificity between the KTA/KSThR and ATA guidelines was 0–1.5% with the original biopsy size criteria, and 0–4.5% with the simulated, similar biopsy size criteria. Although the difference in specificity between the ACR and KTA/KSThR, or ATA guidelines was 39.4–40.9% in dataset 1, and 40.9% in dataset 2 with the original biopsy criteria, the difference in specificity was decreased with the four simulated biopsy criteria to 3.6–10.9% in dataset 1, and 2.5–10.5% in dataset 2.

Unnecessary Biopsy Rate

The estimated unnecessary FNA rate with the original biopsy size criteria was similar between the KTA/KSThR and ATA guidelines in both datasets (dataset 1, 56.9% vs. 55.7%; dataset 2, 59.6%, vs. 59.6%, respectively). However, the unnecessary FNA rate based on the ACR guidelines was significantly lower than the KTA/KSThR or ATA guidelines, in both datasets (dataset 1, 25.3%; dataset 2, 25.8%; all, $p < 0.001$) (Tables 6, 7).

When the original biopsy criteria were applied, the difference in unnecessary FNA rates between the ACR and KTA/KSThR or ATA guidelines was 30.4–31.6 % in dataset 1, and 33.8% in dataset 2; while the difference between the KTA/KSThR and ATA guidelines was 0–1.2% in both datasets. However, when the simulated FNA size criteria were applied, the difference in unnecessary FNA rate between the ACR and KTA/KSThR or ATA guidelines decreased to 2.0–8.7%. The difference between the KTA/KSThR and ATA guidelines was 0–3.8% in both datasets with the four simulated biopsy criteria. Compared to the KTA/KSThR guidelines, the ACR guidelines resulted in significantly lower unnecessary FNA rates ($p \leq 0.008$), with all size criteria for both datasets, except for the simulated biopsy criterion 4 in dataset 2 ($p = 0.193$).

DISCUSSION

Our study results demonstrated high concordance rates (84.1–100%) for nodules of intermediate- or moderate-suspicion, and high- or highly-suspicious categories, between the three guidelines, while the concordance rates for other categories were relatively low and variable. This study also demonstrated that the diagnostic performance of biopsy criteria was similar between the three guidelines at each simulated, similar biopsy size criterion. Additionally, the diagnostic performance resulted in a less than 6% difference in sensitivity and 9% difference in unnecessary FNA rate, in both datasets with different study populations.

 https://doi.org/10.3348/kjr.2019.0215

These results differ from those of several previous studies reporting higher sensitivity and higher unnecessary FNA rates associated with the KTA/KSThR and ATA guidelines, compared to the ACR guidelines (8-10). The discrepancy in these results suggests that the difference in diagnostic performance between the three guidelines mostly depends on the biopsy size criteria, rather than on a difference in the US categorization system for thyroid nodules. Furthermore, the results from the present study were consistent across both datasets, even though they had significantly different malignancy rates and distribution of malignant tumor histological types. These results suggest that the impact of biopsy size thresholds on the diagnostic performance of nodule detection may be applied to different study populations.

Simulated data calculated with similar biopsy size criteria showed that the ACR guidelines resulted in a slightly lower unnecessary FNA rate than the KTA/KSThR and ATA guidelines. This may be explained by the differences between the guidelines in the US categorization system and biopsy indication. The not-suspicious category based on the ACR guidelines included 23.2–26.3% of the low-suspicion nodules based on the KTA/KSThR guidelines, and 65.4–67.8% of the very-low-, and 2.0–10.3% of the low-suspicion nodules based on the ATA guidelines. The vastly different categorization system in the ACR guidelines led to a lower chance of biopsy for benign nodules, compared to the KTA/KSThR and ATA guidelines.

Regardless of the guidelines, the calculated sensitivity of biopsy criteria was lower in dataset 1 than in dataset 2. This outcome is due to a higher proportion of malignant tumors demonstrating low-risk US patterns in dataset 1, compared to dataset 2, and may be related to the difference in the proportion of PTC and FTC between the two datasets. Since the majority of PTC display high- or intermediate-suspicion US patterns (18), while most FTC display intermediate- or low-suspicion US patterns (18, 19), a higher proportion of malignant tumors may be designated to undergo FNA at smaller size thresholds in study populations with a higher proportion of PTC. Therefore, the calculated sensitivity would be affected by a subset of histological types among the malignant tumors in the study population. Another factor influencing the calculated sensitivity could be the size distribution of nodules in the study population. The calculated sensitivity of FNA criteria may be overestimated if the proportion of smaller malignant tumors is relatively low in the study population, or when higher FNA size criteria are applied in clinical practice to study populations. In our study population, most thyroid nodules ≥ 1 cm, except for nodules with benign US patterns, had undergone FNA during the study period. This could minimize the selection bias of the target population with thyroid nodules ≥ 1 cm.

The ACR guidelines adopt a strategy of higher biopsy size threshold and US monitoring for thyroid nodules. This could markedly reduce the unnecessary biopsy rate during the initial evaluation of thyroid nodules (8). Although the sensitivity of this combined approach (biopsy and US monitoring) for malignancy may be sufficiently high (more than 90%), the efficacy of US monitoring for relatively small nodules (1–1.5 or 2.5 cm) may need to be prospectively investigated. It is still uncertain whether growing primary thyroid cancer always precedes nodal (rarely distant) metastasis, or whether US monitoring of tumor growth can effectively prevent the potential risk of increased morbidity related to nodal metastasis or clinically significant local invasion, since small aggressive papillary thyroid microcarcinoma may show gross ETE and macroscopic nodal metastasis in rare cases. According to a recent report from Japan (20), most novel lymph node metastases (11 of 12 patients) were detected during active surveillance, without detection of primary tumor growth. Although the incidence is rare, other concerns include the potential risk of a more aggressive, high-grade malignancy, and possible anaplastic transformation of well-differentiated thyroid cancer (21). The appropriate size criteria for biopsies remain controversial. The ultimate goal of a diagnostic strategy for thyroid nodules is to provide an individualized benefit to patients based on benefit and risk balance, as well as a reduction in the number of unnecessary biopsies. Since it is difficult to simultaneously achieve high-sensitivity and low unnecessary FNA, guidelines should appropriately balance the two for detection of thyroid malignancies.

Our study had several limitations. First, we only included thyroid nodules that had undergone US-guided FNA or CNB; therefore, selection bias might be unavoidable. Second, ETE was not evaluated in this study, which might affect the diagnostic performance of the ATA and ACR guidelines. Third, the estimated diagnostic performance of the ATA guidelines in our study is slightly different from that reported by previous studies (8-11). This is because we categorized unclassified nodules based on the ATA guidelines as intermediate-suspicion nodules, to assess diagnostic performance in real practice. Fourth, we were unable to provide interobserver agreement of US classified

https://doi.org/10.3348/kjr.2019.0215

Korean Journal of Radiology
KJR

nodules in the current study, because we analyzed a previously published database containing two datasets (10, 11). Last, the reference standard for benign diagnosis was FNA cytology or CNB biopsy, as well as surgery, which may inevitably lead to false-negative results.

In conclusion, regardless of the guidelines, the concordance rate of high- or intermediate-suspicion nodules was high, and the diagnostic performance of biopsy size criteria was similar at each simulated, similar biopsy size criterion. The difference between the three guidelines in the diagnostic performance of biopsy criteria for malignancy is mainly influenced by different size thresholds for biopsies, and partly by different US categorization systems for thyroid nodules.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

ORCID iDs

Dong Gyu Na
https://orcid.org/0000-0001-6422-1652
Younghee Yim
https://orcid.org/0000-0002-4224-7832
Eun Ju Ha
https://orcid.org/0000-0002-1234-2919
Jung Hwan Baek
https://orcid.org/0000-0003-0480-4754
Jin Yong Sung
https://orcid.org/0000-0002-8163-4624
Ji-hoon Kim
https://orcid.org/0000-0002-6349-6950
Won-Jin Moon
https://orcid.org/0000-0002-8925-7376

REFERENCES

1. Perros P, Boelaert K, Colley S, Evans C, Evans RM, Gerrard Ba G, et al.; British Thyroid Association. Guidelines for the management of thyroid cancer. *Clin Endocrinol (Oxf)* 2014;81 Suppl 1:1-122
2. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 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* 2016;26:1-133
3. Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L, et al. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules–2016 update. *Endocrine Practice* 2016;22(Supple 1):1-60
4. Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al.; Korean Society of Thyroid Radiology (KSThR) and Korean Society of Radiology. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology consensus statement and recommendations. *Korean J Radiol* 2016;17:370-395
5. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, 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
6. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *Eur Thyroid J* 2017;6:225-237
7. Ha EJ, Baek JH, Na DG. Risk stratification of thyroid nodules on ultrasonography: current status and perspectives. *Thyroid* 2017;27:1463-1468
8. Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, et al. 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* 2018;210:1148-1154
9. Ha EJ, Na DG, Baek JH, Sung JY, Kim JH, Kang SY. US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. *Radiology* 2018;287:893-900
10. Ha EJ, Na DG, Moon WJ, Lee YH, Choi N. Diagnostic performance of ultrasound-based risk stratification systems for thyroid nodules: comparison of the 2015 ATA guidelines with the 2016 KTA/KSThR and 2017 ACR guidelines. *Thyroid* 2018;28:1532-1537
11. Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, 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
12. Ha SM, Baek JH, Na DG, Suh CH, Chung SR, Choi YJ, et al. Diagnostic performance of practice guidelines for thyroid nodules: thyroid nodule size versus biopsy rates. *Radiology* 2019;291:92-99
13. Na DG, Baek JH, Sung JY, Kim JH, Kim JK, Choi YJ, et al. Thyroid imaging reporting and data system risk stratification of thyroid nodules: categorization based on solidity and echogenicity. *Thyroid* 2016;26:562-572
14. Ha EJ, Moon WJ, Na DG, Lee YH, Choi N, Kim SJ, et al. A multicenter prospective validation study for the Korean thyroid imaging reporting and data system in patients with thyroid nodules. *Korean J Radiol* 2016;17:811-821
15. Cibas ES, Ali SZ. The Bethesda system for reporting thyroid cytopathology. *Am J Clin Pathol* 2009;132:658-665

 https://doi.org/10.3348/kjr.2019.0215

16. Jung CK, Min HS, Park HJ, Song DE, Kim JH, Park SY, et al.; Korean Endocrine Pathology Thyroid Core Needle Biopsy Study Group. Pathology reporting of thyroid core needle biopsy: a proposal of the Korean Endocrine Pathology Thyroid Core Needle Biopsy Study Group. *J Pathol Transl Med* 2015;49:288-299
17. Yoon JH, Lee HS, Kim EK, Moon HJ, Kwak JY. Malignancy risk stratification of thyroid nodules: comparison between the thyroid imaging reporting and data system and the 2014 American Thyroid Association management guidelines. *Radiology* 2016;278:917-924
18. Hong MJ, Na DG, Baek JH, Sung JY, Kim JH. Impact of nodule size on malignancy risk differs according to the ultrasonography pattern of thyroid nodules. *Korean J Radiol* 2018;19:534-541
19. Park JW, Kim DW, Kim D, Baek JW, Lee YJ, Baek HJ. Korean thyroid imaging reporting and data system features of follicular thyroid adenoma and carcinoma: a single-center study. *Ultrasonography* 2017;36:349-354
20. Ito Y, Miyauchi A, Oda H. Low-risk papillary microcarcinoma of the thyroid: a review of active surveillance trials. *Eur J Surg Oncol* 2018;44:307-315
21. Han JM, Kim WB, Kim TY, Ryu JS, Gong G, Hong SJ, et al. Time trend in tumour size and characteristics of anaplastic thyroid carcinoma. *Clin Endocrinol (Oxf)* 2012;77:459-464

https://doi.org/10.3348/kjr.2019.0215