EnM
ENDOCRINOLOGY AND METABOLISM

Original Article

Endocrinol Metab 2023;38:117-128
https://doi.org/10.3803/EnM.2023.1670
pISSN 2093-596X · eISSN 2093-5978

Diagnostic Performance of Ultrasound-Based Risk Stratification Systems for Thyroid Nodules: A Systematic Review and Meta-Analysis

Leehi Joo[1], Min Kyoung Lee[2], Ji Ye Lee[3], Eun Ju Ha[4], Dong Gyu Na[5]

[1]Department of Radiology, Korea University Guro Hospital; [2]Department of Radiology, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea; [3]Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul; [4]Department of Radiology, Ajou University Hospital, Ajou University School of Medicine, Suwon; [5]Department of Radiology, Gangneung Asan Hospital, University of Ulsan College of Medicine, Gangneung, Korea

Background: This study investigated the diagnostic performance of biopsy criteria in four society ultrasonography risk stratification systems (RSSs) for thyroid nodules, including the 2021 Korean (K)-Thyroid Imaging Reporting and Data System (TIRADS).

Methods: The Ovid-MEDLINE, Embase, Cochrane, and KoreaMed databases were searched and a manual search was conducted to identify original articles investigating the diagnostic performance of biopsy criteria for thyroid nodules (≥1 cm) in four widely used society RSSs.

Results: Eleven articles were included. The pooled sensitivity and specificity were 82% (95% confidence interval [CI], 74% to 87%) and 60% (95% CI, 52% to 67%) for the American College of Radiology (ACR)-TIRADS, 89% (95% CI, 85% to 93%) and 34% (95% CI, 26% to 42%) for the American Thyroid Association (ATA) system, 88% (95% CI, 81% to 92%) and 42% (95% CI, 22% to 67%) for the European (EU)-TIRADS, and 96% (95% CI, 94% to 97%) and 21% (95% CI, 17% to 25%) for the 2016 K-TIRADS. The sensitivity and specificity were 76% (95% CI, 74% to 79%) and 50% (95% CI, 49% to 52%) for the 2021 K-TIRADS$_{1.5}$ (1.5-cm size cut-off for intermediate-suspicion nodules). The pooled unnecessary biopsy rates of the ACR-TIRADS, ATA system, EU-TIRADS, and 2016 K-TIRADS were 41% (95% CI, 32% to 49%), 65% (95% CI, 56% to 74%), 68% (95% CI, 60% to 75%), and 79% (95% CI, 74% to 83%), respectively. The unnecessary biopsy rate was 50% (95% CI, 47% to 53%) for the 2021 K-TIRADS$_{1.5}$.

Conclusion: The unnecessary biopsy rate of the 2021 K-TIRADS$_{1.5}$ was substantially lower than that of the 2016 K-TIRADS and comparable to that of the ACR-TIRADS. The 2021 K-TIRADS may help reduce potential harm due to unnecessary biopsies.

Keywords: Thyroid nodule; Thyroid neoplasms; Ultrasonography; Biopsy; Meta-analysis

INTRODUCTION

The management of thyroid nodules has become a topic of debate worldwide with the increasing incidence of thyroid carcinomas and increasing number of thyroid incidentalomas [1-3]. Ultrasonography (US) is the standard imaging modality for

Received: 1 February 2023, **Revised:** 11 February 2023,
Accepted: 12 February 2023

Corresponding author: Min Kyoung Lee
Department of Radiology, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 10 63-ro, Yeongdeungpo-gu, Seoul 07345, Korea
Tel: +82-2-3779-2411, **Fax:** +82-2-783-5288, **E-mail:** tosky333mk@gmail.com

evaluating thyroid nodules, and many professional societies have proposed US-based risk stratification systems (RSSs) or Thyroid Imaging Reporting and Data Systems (TIRADSs) for thyroid nodules [4-11]. Though these systems may share the purpose of optimally discriminating malignancy based on US findings, they have different structures for the risk stratification of nodules (pattern-based or point-based systems) and different size cut-offs for biopsy. RSSs are used for triage to select patients for US-guided biopsy and to rule out thyroid malignancy. As triage tests, RSSs play a role in reducing unnecessary nodule biopsies and require an appropriate sensitivity for thyroid malignancy [12]. Therefore, although many studies have evaluated the diagnostic performance of various RSSs or TIRADSs by using the thresholds for classifying nodules into categories [13,14], the diagnostic performance in real-world practice needs to be assessed using the biopsy criteria of each RSS or TIRADS.

A tendency for overdiagnosis leading to overtreatment has been noted in recent years, and the need to reduce the unnecessary biopsy rate is increasingly emphasized. Therefore, many studies have evaluated the unnecessary biopsy rate as an important index in diagnostic performance [15-19]. The recently updated 2021 K-TIRADS [11] raised the size cut-offs for biopsy for low and intermediate-suspicion nodules to reduce the unnecessary biopsy rate because previous studies had shown that the 2016 K-TIRADS afforded notably high sensitivity for malignancy, but had a high rate of unnecessary biopsies [15-17,20,21].

Therefore, this study aimed to evaluate the diagnostic performance of biopsy criteria in four widely used society RSSs, including the American College of Radiology (ACR)-TIRADS, the American Thyroid Association (ATA) system, the European (EU)-TIRADS, and the 2016/2021 Korean (K)-TIRADS.

METHODS

This systematic review and meta-analysis followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [22].

Literature search strategy

A systematic literature search was done through the Ovid-MEDLINE, Embase, Cochrane, and KoreaMed databases with the following search terms until September 7, 2022: [(thyroid)] AND [(cancer) OR (carcinoma) OR (tumor) OR (neoplasm)] AND [(ultrasonography) OR (sonography) OR (ultrasonic) OR (ultrasound)] AND [(screen) OR (detect) OR (early diagnosis) OR (sensitivity) OR (specificity)]. We included studies published in English. Two thyroid radiologists (L.J. and M.K.L.), each with 8 and 9 years of experience, independently searched the literature and selected relevant articles. Any cases of disagreement were solved by consensus after discussion with a third reviewer (D.G.N.) with 23 years of experience.

Inclusion and exclusion criteria

The inclusion criteria were as follows: (1) population: adult patients who underwent thyroid US and had thyroid nodules larger than 1 cm; (2) index test: US RSSs (ACR-TIRADS [9], ATA system [6], EU-TIRADS [8], 2016 K-TIRADS [7], and 2021 K-TIRADS [11]); (3) reference standard: cytopathologic diagnosis (fine-needle aspiration, core needle biopsy, or surgery) with or without imaging follow-up; (4) outcomes: sensitivity, specificity, and unnecessary biopsy rate; and (5) study design: all observational (retrospective or prospective) original articles.

The exclusion criteria were as follows: (1) studies that did not use RSSs; (2) studies without sufficient data to calculate the diagnostic performance for nodules (≥1 cm) based on the estimated true-positive, true-negative, false-positive, and false-negative rates, according to any of the ACR-TIRADS, ATA system, EU-TIRADS, 2016 K-TIRADS, and 2021 K-TIRADS; (3) the presence of a further size limitation for inclusion other than ≥1 cm; (4) studies with a suspected overlapping population or data (in the case of overlap, the study with the larger cohort was included); (5) review articles, case reports, review articles, editorials, letters, and conference abstracts; (6) studies for which the full text was not available in English.

Data extraction

A structured form was used to extract the following information: (1) study characteristics: first author, year of publication, country where each study was performed, study design (prospective/retrospective; single/multicenter), study period, and reference standard; (2) demographic and clinical characteristics: numbers of total and male patients, mean age and range of included patients, numbers of total and malignant nodules, mean size and range of the included nodules; (3) RSSs (ACR-TIRADS, ATA system, EU-TIRADS, 2016 K-TIRADS, and 2021 K-TIRADS); and (4) outcomes: diagnostic performance of biopsy criteria in RSSs, including sensitivity, specificity, and the unnecessary biopsy rate. For the 2021 K-TIRADS, with a range of cut-off sizes for biopsy of intermediate-suspicion nodules (1.0 to 1.5 cm), the diagnostic performance of biopsy criteria was recorded separately as 2021 K-TIRADS$_{1.0}$ and 2021 K-TIRADS$_{1.5}$. The unnecessary biopsy rate was defined as the proportion of biopsy-confirmed benign nodules (false-positive) among all benign nodules (false-positive+true-negative), which also can be calculated as 1-specificity.

Quality assessment

Two reviewers (L.J. and M.K.L.) with 8 and 9 years of experience in thyroid radiology independently extracted the data and performed quality assessment. The quality of included studies was evaluated using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [23]. Any disagreement was solved by consensus after discussion with a third reviewer (D. G.N.) with 23 years of experience.

Data synthesis and analysis

The primary outcome of this meta-analysis was the diagnostic performance of each US RSS for thyroid nodules. Using random-effects modeling, the pooled sensitivity and specificity with 95% confidence intervals (CIs) were evaluated from individual studies. Hierarchical summary receiver operating characteristic (HSROC) curves with 95% CIs and prediction regions were graphically visualized. Publication bias was evaluated using a Deeks' funnel plot, and Deeks' asymmetry test was used to evaluate the *P* value and statistical significance [24]. A secondary outcome was the unnecessary biopsy rate, which was defined as the proportion of biopsy-confirmed benign nodules among all benign nodules. For meta-analytic pooling of the unnecessary biopsy rate, the inverse variance method was used to calculate weights, and their 95% CIs were obtained using DerSimonian-Laird random-effects modeling [25]. The Higgins I^2 statistic was used to determine the heterogeneity (I^2=0% to 40%, insignificant heterogeneity; 30% to 60%, moderate heterogeneity; 50% to 90%, substantial heterogeneity; and 75% to 100%, considerable heterogeneity) [26].

The presence of a threshold effect caused by heterogeneity was visually assessed by the coupled forest plots of pooled sensitivity and specificity. In addition, the threshold effect, which is a positive correlation between sensitivity and the false-positive rate, was calculated; a Spearman correlation coefficient >0.6 between the sensitivity and false-positive rates was considered to indicate a threshold effect [27].

All statistical analyses were performed using STATA version 17.0 (Stata Corp, College Station, TX, USA).

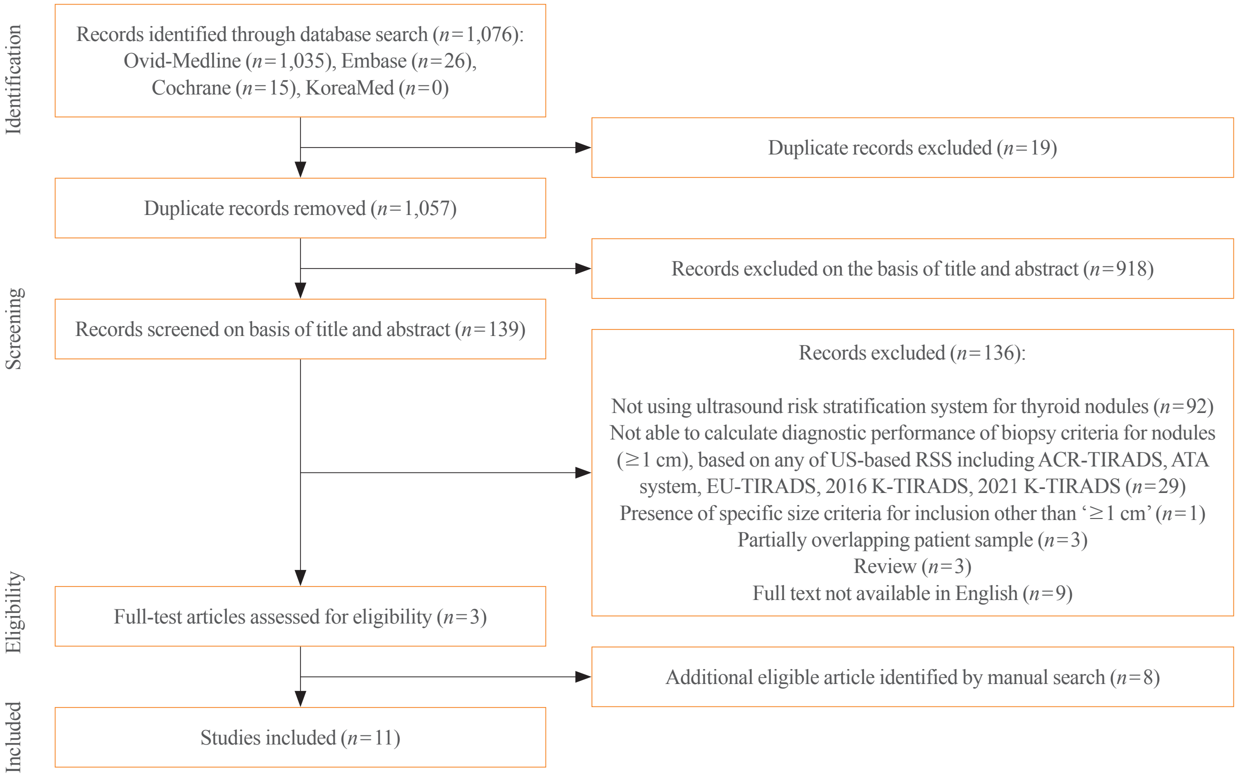
Identification
Records identified through database search (n=1,076): Ovid-Medline (n=1,035), Embase (n=26), Cochrane (n=15), KoreaMed (n=0)
Duplicate records excluded (n=19)
Duplicate records removed (n=1,057)
Records excluded on the basis of title and abstract (n=918)
Screening
Records screened on basis of title and abstract (n=139)
Records excluded (n=136):
Not using ultrasound risk stratification system for thyroid nodules (n=92)
Not able to calculate diagnostic performance of biopsy criteria for nodules (≥1 cm), based on any of US-based RSS including ACR-TIRADS, ATA system, EU-TIRADS, 2016 K-TIRADS, 2021 K-TIRADS (n=29)
Presence of specific size criteria for inclusion other than '≥1 cm' (n=1)
Partially overlapping patient sample (n=3)
Review (n=3)
Full text not available in English (n=9)
Eligibility
Full-test articles assessed for eligibility (n=3)
Additional eligible article identified by manual search (n=8)
Included
Studies included (n=11)

Fig. 1. Flow chart of the selection process. US, ultrasound; ACR, American College of Radiology; TIRADS, Thyroid Imaging Reporting and Data System; ATA, American Thyroid Association; EU, European; K, Korean.

Table 1. Characteristics of the Included Studies

Study	Country	Study design	Period	No. of included patients	Mean age (range), yr	No. of male patients (male proportion, %)	No. of nodules (≥1 cm) diagnosed with the reference standard	No. of malignant nodules (≥1 cm, malignancy rate, %)	Mean size (range), cm	US-based RSSs					Reference standard		
										ACR	ATA	EU	2016 K	2021 K	Surgery	Biopsy (FNA w/wo CNB)	US follow-up[a]
Chung et al. (2021) [33][b]	Korea	Multicenter, retrospective	06/2015–09/2015	5,081	53.2 (19–76)	905 (17.8)	5,708	1,111 (19.5)	2.1 (1–10)	N	N	N	Y	Y[c]	Y	Y	N
Eidt et al. (2023) [32]	Brazil	Single, prospective	01/2019–12/2021	149	55.0 (NA)	20 (13.4)	168	11 (6.5)	2.6 (1.9–4.0)[d]	Y	N	Y	N	N	Y	Y	N
Grani et al. (2019) [16]	Italy	Single, prospective	11/2015–05/2018	477	55.9 (NA)	119 (24.9)	502	36 (7.2)	NA	Y	Y	Y	Y	N	Y	Y	N
Ha et al. (2018) [20]	Korea	Multicenter, retrospective	06/2013–05/2015	750	49.2 (9–81)[e]	156 (20.8)	586[f]	101 (17.2)[f]	1.5 (NA)	Y	Y	N	Y	N	Y	Y	N
Ha et al. (2018) [15]	Korea	Multicenter, retrospective	01/2010–05/2011	1,802	51.2 (13–79)	415 (23.0)	2,000	454 (22.7)	2.0 (1–NA)	Y	Y	N	Y	N	Y	Y	Y
Ha et al. (2021) [29][b]	Korea	Multicenter, retrospective	06/2015–09/2015	5,081	53.2 (19–93)	905 (17.8)	5,708	1,111 (19.5)	2.1 (1–10)	Y	N	Y	N	Y[g]	Y	Y	N
Ha et al. (2019) [17]	Korea	Single, retrospective	01/2013–12/2013	3,190	53.4 (14–94)	673 (21.1)	3,323	856 (25.8)	1.4 (0.3–9.6)	Y	Y	N	Y	N	Y	Y	Y
Huh et al. (2021) [30]	Korea	Single, retrospective	03/2017–01/2019	2,084	50.4 (19–92)	433 (20.8)	2,106	522 (24.8)	2.3 (1–10)	Y	N	N	N	N	Y	Y	N
Middleton et al. (2018) [21]	USA	Multicenter, retrospective	01/2006–12/2010	3,315	54.4 (18–97)	NA	3,179[h]	288 (9.1)[f]	NA	Y	Y	N	Y	N	Y	Y	N
Na et al. (2021) [31]	Korea	Single, retrospective	01/2011–12/2019	3,088	56.0 (47–64)[h]	591 (19.1)	3,826	549 (14.3)	1.7 (1–10)[c]	Y	Y	Y	Y	N	Y	Y	N
Tan et al. (2020) [28]	Malaysia	Single, retrospective	08/2017–01/2020	128	51.8 (NA)	21 (16.7)	144	7 (4.9)	2.1 (NA)	Y	N	Y	Y	N	N	Y	N

RSS, risk stratification system; ACR, American College of Radiology; ATA, American Thyroid Association; EU, European; K, Korean; FNA, fine-needle aspiration; w/wo, with or without; CNB, core needle biopsy; US, ultrasonography; NA, not available.

[a] In studies that regarded follow-up as a reference standard, thyroid nodules with initial benign results on biopsy and decreased or stable size on follow-up US at more than 12 months were finally diagnosed as benign; [b] These two studies shared the same study cohort with different purposes (2021 Korean Thyroid Imaging Reporting and Data System [K-TIRADS] vs. 2016 K-TIRADS [33]; 2021 K-TIRADS vs. foreign RSSs [29]). We representatively cited Chung et al. [33] for the outcomes from the 2021 K-TIRADS (2021 K-TIRADS$_{1.0}$ or 2021 K-TIRADS$_{1.5}$) throughout this study, unless it was necessary to specifically cite the study of Ha et al. [29]; [c] Modified K-TIRADS 1 and modified K-TIRADS 3 correspond to the 2021 K-TIRADS$_{1.0}$ and 2021 K-TIRADS$_{1.5}$, respectively; [d] Median size (interquartile range, cm); [e] Not specified in the article whether M is the mean or median age; [f] These numbers only represent those of thyroid nodules measuring ≥1 cm in each study, after excluding sub-centimeter nodules that were originally included in these study cohorts; [g] Modified K-TIRADS 1.0 cm and modified K-TIRADS 1.5 cm correspond to 2021 K-TIRADS$_{1.0}$ and 2021 K-TIRADS$_{1.5}$, respectively; [h] Median age (interquartile range, years).

RESULTS

Literature search and eligibility criteria

A flow diagram describing study selection is presented in Fig. 1. A total of 1,076 studies were initially identified. Nineteen duplicated studies were excluded, and 918 studies screened on the basis of titles and abstracts were excluded. Afterward, 139 full text articles with potential eligibility were assessed, and 136 studies were further excluded because they did not use any RSS for thyroid nodules (*n*=92), did not provide sufficient data for calculating the diagnostic performance of biopsy criteria for thyroid nodules (≥1 cm) according to any of the RSSs (ACR-TIRADS, ATA system, EU-TIRADS, 2016 K-TIRADS, or 2021 K-TIRADS) (*n*=29), had a further specific size limitation among thyroid nodules (≥1 cm) (*n*=1), were suspected of having an overlapping study population (*n*=3), were reviews (*n*=3), and were not written in English (*n*=9). Eight studies [16,17,21,28-32] were added after searching the bibliographies of these articles. Finally, a total of 11 articles were included [15-17,20,21,28-33].

Characteristics of the included studies

The characteristics of the 11 included studies are summarized in Table 1. Three studies [16,28,32] were prospectively designed, and five [15,20,21,29,33] were multicenter studies. The number of included patients ranged from 128 to 5,081 in all 11 studies, with the proportion of male patients ranging from 13.4% to 24.9% in 10 studies, excluding that of Middleton et al. [21], in which the data were not available. The mean or median age of the included patients ranged from 49.2 to 56 years, except for one article without age information. The number of included nodules ranged from 144 to 5,708, with the proportion of malignant nodules ranging from 6.5% to 29.5%. Diagnostic performance with respect to the biopsy criteria was reported with the following distribution: ACR (*n*=10) [15-17,20,21,28-32], ATA (*n*=6) [15-17,20,21,31], EU-TIRADS (*n*=5) [16,28,29,31,32], 2016 K-TIRADS (*n*=8) [15-17,20,21,28,31,33], and 2021 K-TIRADS (2021 K-TIRADS$_{1.0}$ and 2021 K-TIRADS$_{1.5}$) (*n*=1) [33]. All studies used both cytologic and histopathologic findings as the reference standard, except one [28] that used cytology as the only reference standard. In two studies [15,17], US follow-up was used as one of the reference standards for benign nodules, and thyroid nodules with initial benign results on biopsy and decreased or stable size on follow-up US after more than 12 months were finally classified as benign nodules.

Quality assessment

Nine studies fulfilled five domains, one study fulfilled four domains, and one study fulfilled all seven domains (Fig. 2). Ten studies [15-17,20,21,28,29,31-33] had a low-risk of bias for patient selection regarding consecutively registered patients. Patient selection was unclear in one study [30]. All studies had a low-risk of bias in the index test domain owing to the use of specified RSSs. One study had a low-risk of bias in the reference standard because they specified that a pathologist was blinded to the radiology report, while the others had an unclear risk of bias [28]. One study had a low-risk of bias in the flow and timing domain because cytology was the only reference standard in the study [28]. The flow and timing domain was unclear in the other 10 studies [15-17,20,21,25,29-33]. All 11 studies were categorized as having low concerns for applicability in the patient selection, index test, and reference standard domains.

Diagnostic performance

The diagnostic performance of biopsy criteria in RSSs is sum-

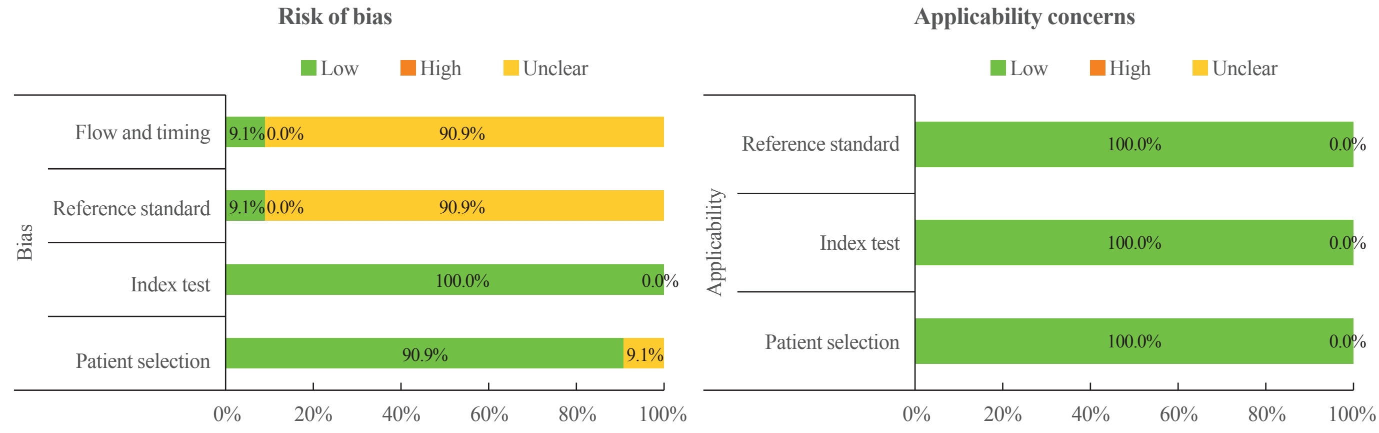

Fig. 2. Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) criteria for the 11 included studies.

Table 2. Diagnostic Performance of Biopsy Criteria in Four Ultrasound-Based Risk Stratification Systems

RSS	Study	No. of included nodules (≥1 cm, malignant/total)	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy	Unnecessary biopsy rate (1-specificity)
ACR-TIRADS	Eidt et al. (2023) [32]	11/168	100 (100–100.0)	28.7 (21.6–35.7)	8.9 (3.9–14.0)	100.0 (100–100.0)	33.3 (26.3–41.0)	71.3
	Grani et al. (2019) [16]	36/502	83.3 (67.2–93.6)	56.2 (51.6–60.8)	12.8 (8.8–17.8)	97.8 (95.2–99.2)	58.2 (53.7–62.5)	43.8
	Ha et al. (2018) [20]	101/586	80.2 (71.1–87.5)	68.9 (64.5–73.0)	34.9 (31.3–38.7)	94.4 (91.8–96.1)	70.8 (67.0–74.5)	31.1
	Ha et al. (2018) [15]	454/2,000	74.7 (70.7–78.7)	67.3 (65.0–69.7)	40.2 (36.9–43.5)	90.1 (88.3–91.8)	69.0 (67.0–71.0)	32.7
	Ha et al. (2021) [29]	1,111/5,708	76.1 (73.5–78.5)	61.8 (60.4–63.2)	32.5 (30.7–34.3)	91.4 (90.4–92.4)	64.6 (63.3–65.8)	38.2 (36.8–39.6)
	Ha et al. (2019) [17]	321/1,938	60.1 (54.5–65.5)	75.2 (73.0–77.3)	32.5 (29.9–35.3)	90.5 (89.2–91.6)	72.7 (70.7–74.7)	24.8
	Huh et al. (2021) [30]	522/2,106	86.4 (83.5–89.3)	63.1 (60.8–65.5)	43.6 (40.6–46.6)	93.4 (91.9–94.9)	68.9 (66.9–70.9)	36.9
	Middleton et al. (2018) [21]	288/3,179	83.3 (78.5–87.5)	49.9 (48.1–52.8)	14.2 (13.5–15.0)	96.8 (95.9–97.5)	53.0 (51.2–54.7)	50.1
	Na et al. (2021) [31]	549/3,826	79.6 (76.0–82.9)	65.2 (63.6–66.9)	27.7 (25.5–30.0)	95.0 (94.0–95.9)	67.3 (65.8–68.8)	34.8
	Tan et al. (2020) [28]	7/144	85.7 (42.7–97.4)	56.2 (47.8–64.2)	9.1 (6.5–12.5)	98.7 (92.6–99.8)	57.6 (49.1–65.8)	43.8
ATA system	Grani et al. (2019) [16]	36/502	75.0 (57.8–87.9)	45.3 (40.7–49.9)	9.6 (6.4–13.6)	95.9 (92.4–98.1)	47.4 (43.0–51.9)	54.7
	Ha et al. (2018) [20]	101/586	95.0 (88.8–98.4)	38.1 (33.8–42.6)	24.2 (22.8–25.8)	97.4 (94.0–98.9)	48.0 (43.8–52.1)	61.9
	Ha et al. (2018) [15]	454/2,000	89.6 (86.9–92.5)	33.2 (30.8–35.5)	28.3 (25.9–30.6)	91.6 (89.3–93.9)	46.0 (43.8–48.2)	66.8
	Ha et al. (2019) [17]	321/1,938	92.5 (89.1–95.2)	34.0 (31.6–36.3)	21.8 (21.0–22.6)	95.8 (93.9–97.1)	43.7 (41.4–45.9)	66.0
	Middleton et al. (2018) [21]	288/3,179	92.7 (89.1–95.4)	17.0 (15.7–18.4)	10.0 (9.7–10.4)	95.9 (93.9–97.3)	23.9 (22.4–25.4)	83.0
	Na et al. (2021) [31]	549/3,826	84.0 (80.6–86.9)	41.6 (39.9–43.3)	19.4 (17.8–21.1)	93.9 (92.6–95.1)	47.7 (46.0–49.2)	58.4
EU-TIRADS	Eidt et al. (2023) [32]	11/168	90.9 (73.9–99.9)	19.1 (13.0–25.3)	7.3 (2.9–11.7)	96.8 (90.6–100.0)	23.8 (17.6–31.0)	80.9
	Grani et al. (2019) [16]	36/502	86.1 (70.5–95.3)	32.0 (27.8–36.4)	8.9 (6.1–12.4)	96.7 (92.6–98.9)	35.9 (31.7–40.2)	68.0
	Ha et al. (2021) [29]	1,111/5,708	84.6 (82.4–86.6)	39.3 (37.9–40.7)	25.2 (23.8–26.6)	91.4 (90.0–92.5)	48.1 (46.8–49.4)	60.7 (59.3–62.1)
	Na et al. (2021) [31]	549/3,826	88.3 (85.4–90.9)	33.4 (31.7–35.0)	18.2 (16.7–19.7)	94.5 (93.0–95.7)	41.2 (39.7–42.8)	66.6
	Tan et al. (2020) [28][a]	7/144	85.7 (42.1–99.6)	38.7 (30.5–47.4)	6.7 (4.9–9.0)	98.2 (89.5–99.7)	41.0 (32.9–49.5)	61.3

(*Continued to the next page*)

Table 2. Continued

RSS	Study	No. of included nodules (≥1 cm, malignant/total)	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy	Unnecessary biopsy rate (1-specificity)
2016 K-TIRADS	Chung et al. (2021) [33]	1,111/5,708	94.9 (93.4–96.0)	24.4 (23.2–25.7)	23.3 (22.1–24.5)	95.2 (93.8–96.3)	38.1 (36.9–39.4)	75.6
	Grani et al. (2019) [16]	36/502	91.7 (77.5–98.2)	17.8 (14.4–21.6)	7.9 (5.5–11)	96.5 (90.2–99.3)	23.1 (19.5–27.1)	82.2
	Ha et al. (2018) [20]	101/586	100 (96.4–100)	28.2 (24.3–32.5)	22.5 (21.5–23.5)	100	40.6 (36.6–44.7)	71.8
	Ha et al. (2018) [15]	454/2,000	94.5 (92.4, 96.6)	26.4 (24.2–28.6)	27.4 (25.2–29.6)	94.2 (92.0–96.4)	41.9 (39.7–44.0)	73.6
	Ha et al. (2021) [29]	321/1,938	93.5 (90.2–95.9)	28.7 (26.5–31.0)	20.6 (20.0–21.4)	95.7 (93.6–97.1)	39.4 (37.2–41.6)	71.3
	Middleton et al. (2018) [21]	288/3,179	96.2 (93.3–98.1)	15.4 (14.1–16.7)	10.2 (9.9–10.4)	97.6 (95.7–98.6)	22.7 (21.2–24.2)	84.6
	Na et al. (2021) [31]	549/3,826	96.9 (95.1–98.2)	18.6 (17.3–20.0)	16.6 (15.4–18.0)	97.3 (95.7–98.4)	29.9 (28.4–31.4)	81.4
	Tan et al. (2020) [28]	7/144	100 (59.0–100)	12.4 (7.4–19.1)	5.5 (5.2–5.9)	100 (100–100.0)	16.7 (11.0–23.8)	87.6
2021 K-TIRADS$_{1.0}$[a]	Chung et al. (2021) [33]	1,111/5,708	91.0 (89.2–92.5)	39.7 (38.3–41.1)	26.7 (25.3–28.2)	94.8 (93.7–95.7)	49.7 (48.4–51.0)	60.3 (58.9–61.7)
2021 K-TIRADS$_{1.5}$[b]	Chung et al. (2021) [33]	1,111/5,708	76.1 (73.6–78.6)	50.2 (48.7–51.6)	27.0 (25.5–28.6)	89.7 (88.5–90.8)	55.2 (53.9–56.5)	49.8 (48.4–51.3)

RSS, risk stratification system; ACR, American College of Radiology; TIRADS, Thyroid Imaging Reporting and Data System; ATA, American Thyroid Association; EU, European; K, Korean.

[a] 2021 K-TIRADS$_{1.0}$ indicates that 1.0 cm was used as a cut-off for intermediate-suspicion nodules; [b] 2021 K-TIRADS$_{1.5}$ indicates that 1.5 cm was used as a cut-off for intermediate-suspicion nodules.

marized in Table 2. Among the studies evaluating the diagnostic performance of biopsy criteria in RSSs, the pooled sensitivity and specificity were 82% (95% CI, 74% to 87%) and 60% (95% CI, 52% to 67%) for the ACR-TIRADS, 89% (95% CI, 85% to 93%) and 34% (95% CI, 26% to 42%) for the ATA system, 88% (95% CI, 81% to 92%) and 42% (95% CI, 22% to 67%) for the EU-TIRADS, and 96% (95% CI, 94% to 97%) and 21% (95% CI, 17% to 25%) for the 2016 K-TIRADS (Fig. 3). A large-population multicenter study of the 2021 K-TIRADS [33] showed sensitivity and specificity of 91% (95% CI, 89% to 93%) and 40% (95% CI, 38% to 41%) with the 1.0-cm cut-off for intermediate-suspicion nodules (2021 K-TIRADS$_{1.0}$), and 76% (95% CI, 74% to 79%) and 50% (95% CI, 49% to 52%) with the 1.5-cm cut-off for intermediate-suspicion nodules (2021 K-TIRADS$_{1.5}$), respectively. All studies showed considerable heterogeneity (I^2>75%), except for the sensitivity of EU-TIRADS (I^2=0%).

Unnecessary biopsy rates

The unnecessary biopsy rate in RSSs is summarized in Table 2. The pooled unnecessary biopsy rates of the ACR-TIRADS, ATA system, EU-TIRADS, and 2016 K-TIRADS were 41% (95% CI, 32% to 49%), 65% (95% CI, 56% to 74%), 68% (95% CI, 60% to 75%), and 79% (95% CI, 74% to 83%), respectively (Fig. 4). All studies showed considerable heterogeneity (I^2>75%). A large-population multicenter study of the 2021 K-TIRADS [33] showed an unnecessary biopsy rate of 60% (95% CI, 59% to 62%) with the 1.0-cm cut-off for intermediate-suspicion nodules (2021 K-TIRADS$_{1.0}$) and 50% (95% CI, 48% to 51%) with the 1.5-cm cut-off for intermediate-suspicion nodules (2021 K-TIRADS$_{1.5}$). The pooled unnecessary biopsy rate in all RSSs was 60% (95% CI, 54% to 67%).

DISCUSSION

Our study, which included 11 studies with 27,250 nodules,

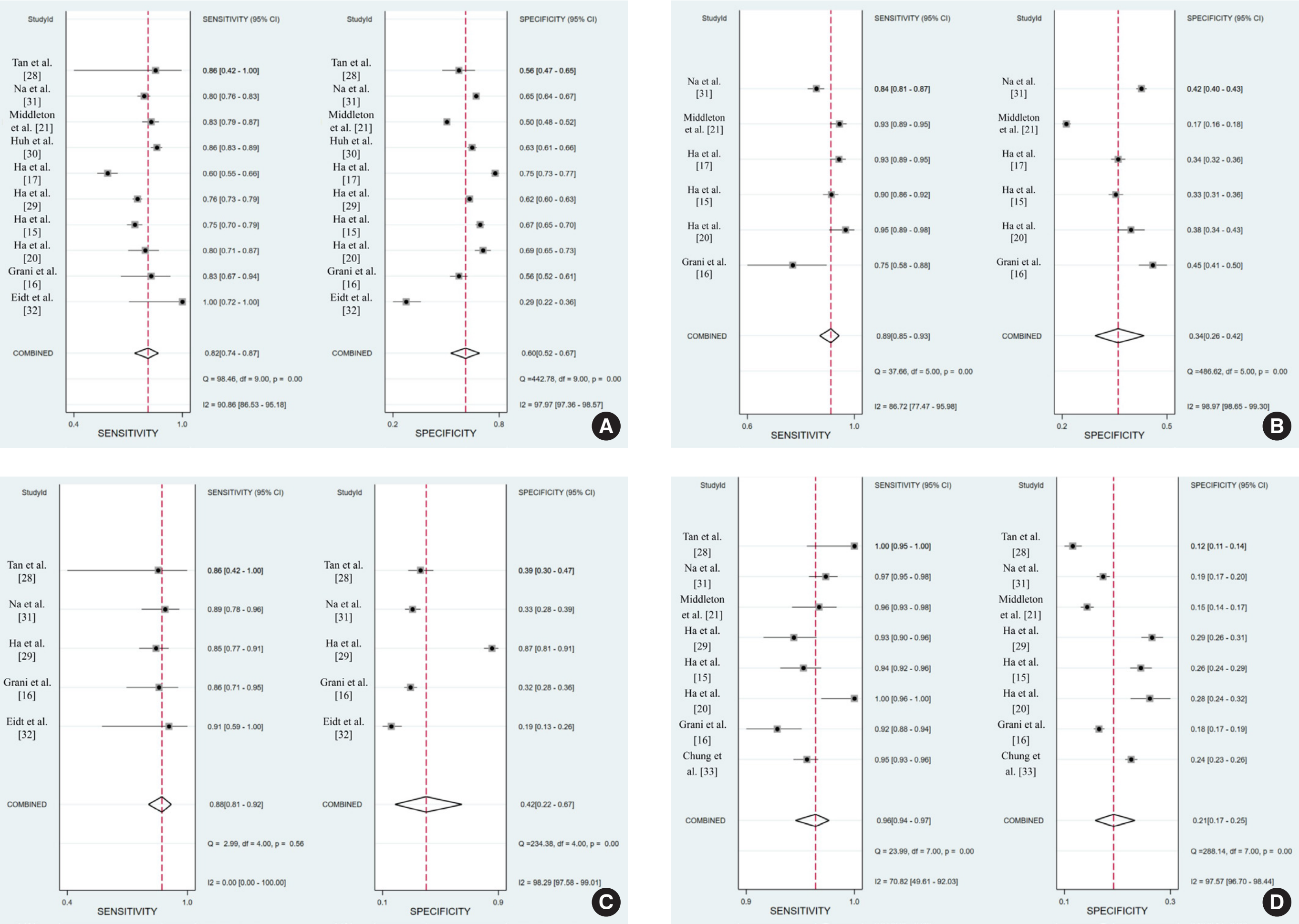

Fig. 3. Sensitivity and specificity of the (A) American College of Radiology (ACR)-Thyroid Imaging Reporting and Data System (TI-RADS), (B) American Thyroid Association (ATA) system, (C) European (EU)-TIRADS, and (D) 2016 Korean (K)-TIRADS. CI, confidence interval.

showed that the diagnostic performance of US-based biopsy criteria was variable among the RSSs, ranging from 76% to 96% for sensitivity, from 21% to 60% for specificity, and from 41% to 79% for the unnecessary biopsy rate. The 2016 K-TIRADS had the highest sensitivity and unnecessary biopsy rate, and the ACR-TIRADS had the lowest sensitivity and unnecessary biopsy rate among the pooled data. The 2021 K-TIRADS$_{1.5}$ had a similar sensitivity and unnecessary biopsy rate compared to those of the ACR-TIRADS, and the 2021 K-TIRADS showed a substantially lower unnecessary biopsy rate with either cut-off (1 or 1.5 cm) for intermediate-suspicion nodules than that of the 2016 K-TIRADS.

In this study, we investigated the diagnostic performance of biopsy criteria in RSSs, including the 2021 K-TIRADS, for clinically relevant thyroid nodules (≥1 cm). Although many studies have investigated the diagnostic performance of RSSs, very few studies have specifically focused on reviewing the diagnostic performance of the biopsy criteria in RSSs [34,35]. In a review article by Castellana et al. [34], diagnostic performance was also variable among RSSs and ranged from 54% to 87% for sensitivity and from 28% to 64% for specificity. Additionally, the tendency for higher sensitivity with the 2016 K-TIRADS (86%; 95% CI, 73% to 94%) and the ATA system (87%; 95% CI, 75% to 94%) and higher specificity with the ACR-TIRADS (74%; 95% CI, 61% to 83%) was similar to the findings of our study. However, there were two essential points that made our study different: (1) the inclusion of the 2021 K-TIRADS; and (2) the exclusion of data from sub-centimeter nodules. We only included studies with relevant data for nodules over 1 cm, as sub-centimeter nodules are not routinely recommended to be biopsied.

The unnecessary biopsy rate has received attention in studies evaluating the diagnostic performances of RSSs [15-19,35] with

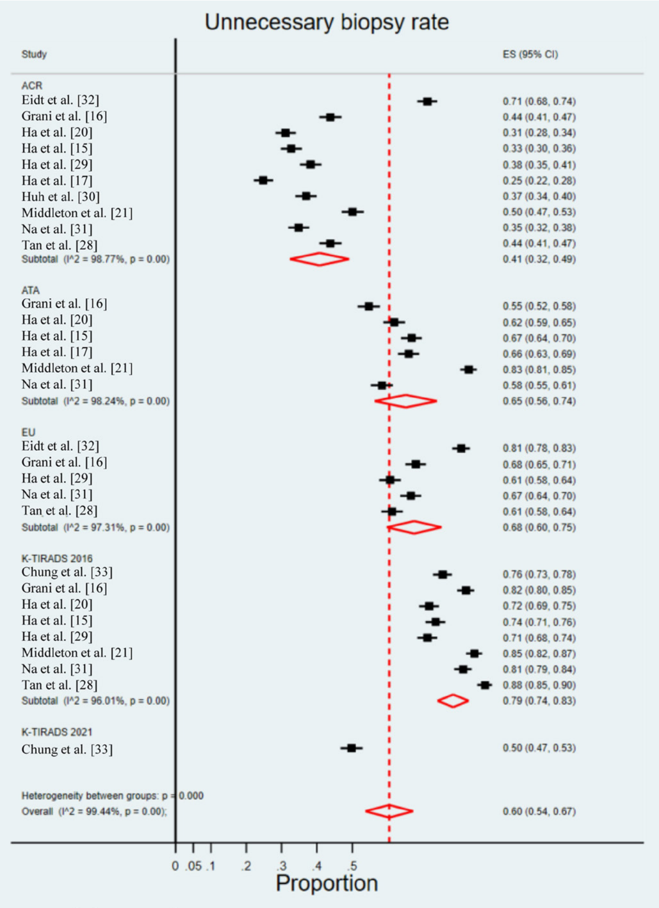

Fig. 4. Unnecessary biopsy rates for the four risk stratification systems. ES, effect size; CI, confidence interval; ACR, American College of Radiology; ATA, American Thyroid Association; EU, European; K-TIRADS, Korean Thyroid Imaging Reporting and Data System.

respect to the potential harm of unnecessary biopsy. False-positive results carry the risk of potential complications and increased costs due to an increased number of biopsies, although US-guided biopsy is a safe procedure, and inconclusive biopsy results may lead to repeated biopsies or unnecessary diagnostic surgery for some nodules [36]. However, there are various definitions of the unnecessary biopsy rate: (1) the percentage of benign nodules among nodules requiring biopsy (1–positive predictive value) [28,33]; (2) the percentage of benign nodules requiring biopsy among all nodules [15,17,28,33]; and (3) the percentage of benign nodules requiring biopsy among all benign nodules (1–specificity) [20,31]. We used the third definition considering the heterogeneity in the prevalence of malignant tumors among the included studies, because the unnecessary biopsy rate defined using the other definitions depends on the prevalence of malignant tumors in the study population. In a review of the unnecessary biopsy rate for thyroid nodules according to four RSSs [35], the first definition of the unnecessary biopsy rate was applied, and the ACR-TIRADS showed a significantly lower unnecessary biopsy rate of 25% (95% CI, 22% to 29%) than that of the ATA system (51%; 95% CI, 44% to 58%; *P*<0.001) and the 2016 K-TIRADS (55%; 95% CI, 42% to 67%; *P*<0.001). In our study, the pooled unnecessary biopsy rate of the 2016 K-TIRADS (79%; 95% CI, 74% to 83%) was also higher than that of the ACR-TIRADS (41%; 95% CI, 33% to 49%) despite a different definition of the unnecessary biopsy rate. However, the unnecessary biopsy rate of the 2021 K-TIRADS$_{1.5}$ was reported to be as low as 50% and was relatively similar to that of ACR-TIRADS [29].

Our study is unique in that it includes the 2021 K-TIRADS. The diagnostic performance of the 2021 K-TIRADS was separately described in this study as 2021 K-TIRADS$_{1.0}$ and 2021 K-TIRADS$_{1.5}$ according to each size cut-off, considering the suggested range of 1 to 1.5 cm for biopsy in intermediate-suspicion nodules in the 2021 K-TIRADS. Since most missed malignancies will be small (<1.5 cm) low-risk tumors, it may be reasonable to apply the size cut-off of 1.5 cm for biopsy in most intermediate-suspicion nodules without high-risk clinical or US features of metastasis or gross extrathyroidal extension despite the risk of decreased sensitivity for malignant tumors. However, we may selectively apply the size cut-off of 1 cm for biopsy in some patients with high-risk factors who require higher sensitivity for malignant tumors [11]. The unnecessary biopsy rate of the 2021 K-TIRADS$_{1.5}$ was lower than those of the 2016 K-TIRADS, EU-TIRADS, and ATA system, but was similar to that of the ACR-TIRADS. According to a study comparing the diagnostic performance of biopsy criteria in RSSs [29], the unnecessary biopsy rate of small thyroid nodules (1 to 2 cm) in the 2021 K-TIRADS$_{1.5}$ was the lowest among RSSs, even compared with the ACR-TIRADS. Accordingly, the difference in the unnecessary biopsy rate between the 2016 and 2021 K-TIRADS$_{1.5}$ is due to the reduced number of unnecessary biopsies in small thyroid nodules (1 to 2 cm) by the 2021 K-TIRADS$_{1.5}$. Although the 2021 K-TIRADS$_{1.5}$ had lower sensitivity than the 2016 K-TIRADS, Chung et al. [33] showed that the decrease of sensitivity was exclusively noted for small thyroid nodules (1 to 2 cm) and demonstrated that most missing malignant tumors would be small low-risk tumors. US surveillance can mitigate the decreased sensitivity for small thyroid nodules (1 to 2 cm) in the 2021 K-TIRADS$_{1.5}$.

This study has limitations to note. First, only one relevant study evaluated the diagnostic performance of biopsy criteria in the 2021 K-TIRADS. Although that study was a multicenter study with a large sample size, further validation studies are needed. Second, studies that did not provide the specific outcomes for nodules (≥1 cm) could not be included according to the eligibility criteria. Third, we only presented pooled sensitivity, specificity, and unnecessary biopsy rates among studies without meta-regression due to the paucity of studies employing the 2021 K-TIRADS. It would be worthwhile to perform meta-regression for comparison between RSSs in the future, after more studies adopt the 2021 K-TIRADS.

In conclusion, the 2021 K-TIRADS showed a substantially lower unnecessary biopsy rate than that of the 2016 K-TIRADS, while maintaining an appropriate diagnostic sensitivity for clinically relevant thyroid malignancy. The 2021 K-TIRADS may help reduce the potential harm due to unnecessary biopsies.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We acknowledge and thank Miyoung Choi (National Evidence-based Healthcare Collaborating Agency, Division of Health Technology Assessment Research) and Chang Hee Cho (The Korean Society of Radiology), who contributed to searching and interpreting the evidence. This study was supported by the Korean Thyroid Association and a research fund from National Cancer Center (grant number: 2112570-3).

AUTHOR CONTRIBUTIONS

Conception or design: M.K.L., D.G.N. Acquisition, analysis, or interpretation of data: L.J., M.K.L., D.G.N. Drafting the work or revising: L.J., M.K.L., J.Y.L., E.J.H., D.G.N. Final approval of the manuscript: M.K.L., D.G.N.

ORCID

Leehi Joo *https://orcid.org/0000-0002-5527-0476*
Min Kyoung Lee *https://orcid.org/0000-0003-3172-3159*

REFERENCES

1. Hoang JK, Grady AT, Nguyen XV. What to do with incidental thyroid nodules identified on imaging studies? Review of current evidence and recommendations. Curr Opin Oncol 2015;27:8-14.
2. Vaccarella S, Dal Maso L, Laversanne M, Bray F, Plummer M, Franceschi S. The impact of diagnostic changes on the rise in thyroid cancer incidence: a population-based study in selected high-resource countries. Thyroid 2015;25:1127-36.
3. Seib CD, Sosa JA. Evolving understanding of the epidemiology of thyroid cancer. Endocrinol Metab Clin North Am 2019;48:23-35.
4. Perros P, Boelaert K, Colley S, Evans C, Evans RM, Gerrard Ba G, et al. Guidelines for the management of thyroid cancer. Clin Endocrinol (Oxf) 2014;81 Suppl 1:1-122.
5. Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedus 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. Endocr Pract 2016;22:622-39.
6. 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.
7. Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology consensus statement and recommendations. Korean J Radiol 2016;17:370-95.
8. 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;6:225-37.
9. 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;14:587-95.
10. Zhou J, Yin L, Wei X, Zhang S, Song Y, Luo B, et al. 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. Endocrine 2020;70:256-79.
11. Ha EJ, Chung SR, Na DG, Ahn HS, Chung J, Lee JY, et al. 2021 Korean thyroid imaging reporting and data system and imaging-based management of thyroid nodules: Korean So-

ciety of Thyroid Radiology consensus statement and recommendations. Korean J Radiol 2021;22:2094-123.

12. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol 2010;63: 883-91.
13. Kim DH, Chung SR, Choi SH, Kim KW. Accuracy of thyroid imaging reporting and data system category 4 or 5 for diagnosing malignancy: a systematic review and meta-analysis. Eur Radiol 2020;30:5611-24.
14. Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH. Diagnostic performance of four ultrasound risk stratification systems: a systematic review and meta-analysis. Thyroid 2020;30:1159-68.
15. 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.
16. 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.
17. 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;291:92-9.
18. Xu T, Wu Y, Wu RX, Zhang YZ, Gu JY, Ye XH, et al. Validation and comparison of three newly-released thyroid imaging reporting and data systems for cancer risk determination. Endocrine 2019;64:299-307.
19. Yoon SJ, Na DG, Gwon HY, Paik W, Kim WJ, Song JS, et al. Similarities and differences between thyroid imaging reporting and data systems. AJR Am J Roentgenol 2019;213: W76-84.
20. 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 American Thyroid Association guidelines with the 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology and 2017 American College of Radiology guidelines. Thyroid 2018; 28:1532-7.
21. 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;210:1148-54.
22. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264-9.
23. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
24. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;58:882-93.
25. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-88.
26. Kim KW, Lee J, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers. Part I: General guidance and tips. Korean J Radiol 2015;16:1175-87.
27. Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002;2:9.
28. Tan L, Tan YS, Tan S. Diagnostic accuracy and ability to reduce unnecessary FNAC: a comparison between four Thyroid Imaging Reporting Data System (TI-RADS) versions. Clin Imaging 2020;65:133-7.
29. Ha EJ, Shin JH, Na DG, Jung SL, Lee YH, Paik W, et al. Comparison of the diagnostic performance of the modified Korean Thyroid Imaging Reporting and Data System for thyroid malignancy with three international guidelines. Ultrasonography 2021;40:594-601.
30. Huh S, Yoon JH, Lee HS, Moon HJ, Park VY, Kwak JY. Comparison of diagnostic performance of the ACR and Kwak TIRADS applying the ACR TIRADS’ size thresholds for FNA. Eur Radiol 2021;31:5243-50.
31. Na DG, Paik W, Cha J, Gwon HY, Kim SY, Yoo RE. Diagnostic performance of the modified Korean Thyroid Imaging Reporting and Data System for thyroid malignancy according to nodule size: a comparison with five society guidelines. Ultrasonography 2021;40:474-85.
32. Eidt LB, Nunes de Oliveira C, Lagos YB, Solera GL, Izquierdo R, Meyer EL, et al. A prospective comparison of ACR-TIRADS and EU-TIRADS in thyroid nodule assessment for FNA-US. Clin Endocrinol (Oxf) 2023;98:415-25.

33. Chung SR, Ahn HS, Choi YJ, Lee JY, Yoo RE, Lee YJ, et al. Diagnostic performance of the modified Korean thyroid imaging reporting and data system for thyroid malignancy: a multicenter validation study. Korean J Radiol 2021;22:1579-86.

34. Castellana M, Castellana C, Treglia G, Giorgino F, Giovanella L, Russ G, et al. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. J Clin Endocrinol Metab 2020;105:dgz170.

35. Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH. Unnecessary thyroid nodule biopsy rates under four ultrasound risk stratification systems: a systematic review and meta-analysis. Eur Radiol 2021;31:2877-85.

36. Ha EJ, Na DG, Baek JH. Korean thyroid imaging reporting and data system: current status, challenges, and future perspectives. Korean J Radiol 2021;22:1569-78.