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

Do Hyun Kim, MD, PhD[1], Sung Won Kim, MD, PhD[1], Mohammed Abdullah Basurrah, MD[2], Jueun Lee, MD[3], Se Hwan Hwang, MD, PhD[3]

Evidence Synthesis and Decision Analysis · Systematic Review/Meta-Analysis

Keywords

biopsy, diagnostic imaging, fine-needle, network meta-analysis, thyroid neoplasms, thyroid nodule

Submitted: Sep 21, 2022
Revision requested: Oct 21, 2022
Revision received: Nov 14, 2022
Accepted: Jan 16, 2023
First published online: Feb 8, 2023
Version of record: Mar 22, 2023

An electronic supplement is available online at doi.org/10.2214/AJR.22.28556.

The authors declare that there are no disclosures relevant to the subject matter of this article.

Supported by the National Research Foundation of Korea (grants 2022R1F1A1066232, 2019M3A9H2032424, 2019M3E5D5064110) and the Ministry of Trade, Industry and Energy (grant 20012378), neither of which had any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

BACKGROUND. Risk stratification systems for evaluating thyroid nodules on ultrasound use varying approaches to classify levels of suspicion for malignancy, leading to variable performance.

OBJECTIVE. The purpose of this study was to perform a network meta-analysis comparing six risk stratification systems used to evaluate thyroid nodules on ultrasound in terms of their diagnostic performance for the detection of thyroid cancer.

EVIDENCE ACQUISITION. Five bibliometric databases were searched for studies published through August 31, 2022, that compared at least two of six ultrasound risk stratification systems (the American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi [AACE/ACE/AME] system; American College of Radiology Thyroid Imaging Reporting and Data System [ACR TI-RADS]; the American Thyroid Association [ATA] risk stratification system; European Thyroid Association Thyroid Imaging Reporting and Data System [EU-TIRADS]; the Korean Thyroid Imaging Reporting and Data System [K-TIRADS] endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology; and the Thyroid Imaging Reporting and Data System developed by Kwak et al. [Kwak TIRADS]) in terms of their diagnostic performance for the detection of thyroid cancer, with cytologic or histologic evaluation used as a reference standard. The studies' risk of bias was evaluated using the Newcastle-Ottawa Scale. A meta-analysis of each system was performed to identify the risk category threshold that had the highest accuracy as well as the highest sensitivity and specificity at this threshold. Network meta-analysis was used to perform hierarchic ranking and identify the systems having the highest sensitivities and specificities at each system's most accurate threshold.

EVIDENCE SYNTHESIS. The analysis included 39 studies with 49,661 patients. All studies were of fair (n = 17) or good (n = 22) quality. The most accurate risk category thresholds were class 3 (high risk) for the AACE/ACE/AME system, TR5 (highly suspicious) for ACR TI-RADS, EU-TIRADS 5 (high risk) for EU-TIRADS, 4c (moderate concern but not classic for malignancy) for Kwak TIRADS, K-TIRADS 5 (high suspicion) for K-TIRADS, and high suspicion for the ATA system. At these thresholds, the systems had sensitivity of 64–77% and specificity of 82–90%. Network meta-analysis identified the highest sensitivity and highest specificity for ACR TI-RADS, followed by K-TIRADS.

CONCLUSION. Of six risk stratification systems, ACR TI-RADS had the highest diagnostic performance for the detection of thyroid nodules on ultrasound.

CLINICAL IMPACT. This network meta-analysis can inform decisions regarding implementation of the risk stratification systems and can aid future system updates.

Thyroid nodules are present in 19–68% of the general population, and 6.7–15% of thyroid nodules are malignant [1–3]. Ultrasound is the preferred imaging modality for thyroid nodule characterization because it is readily available, noninvasive, and cost-effective [4]. Therefore, guidelines recommend ultrasound-based management of thyroid nodules [1, 5, 6], and risk stratification methods have been developed to standardize ultrasound evaluation and determine the need for ultrasound-guided fine-needle aspiration (FNA). The most commonly used risk stratification systems for thyroid nodules are

ARRS is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide continuing medical education activities for physicians.

The ARRS designates this journal-based CME activity for a maximum of 1.00 AMA PRA Category 1 Credit™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

To access the article for credit, follow the prompts associated with the online version of this article.

doi.org/10.2214/AJR.22.28556

AJR 2023; 220:791–804

ISSN-L 0361–803X/23/2206–791

[1]Department of Otolaryngology–Head and Neck Surgery, Seoul Saint Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea.

[2]Department of Surgery, College of Medicine, Taif University, Taif, Saudi Arabia.

[3]Department of Otolaryngology–Head and Neck Surgery, Bucheon Saint Mary's Hospital, College of Medicine, The Catholic University of Korea, 327 Sosa-ro, Bucheon-si, Gyeonggi-do 14647, Korea. Address correspondence to S. H. Hwang (yellobird@catholic.ac.kr).

the American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi (AACE/ACE/AME) system, the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS), the American Thyroid Association (ATA) classification, the European Thyroid Association Thyroid Imaging Reporting and Data System (EU-TIRADS), the Korean Thyroid Imaging Reporting and Data System (K-TIRADS) endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology, and the Thyroid Imaging Reporting and Data System developed by Kwak et al. (Kwak TIRADS) [7]. These guidelines provide algorithms for deciding which nodules warrant FNA versus ultrasound follow-up or no further evaluation. However, the approach for classification of suspicious ultrasound features as well as the threshold nodule sizes for determining the need for FNA vary among the systems, confounding data interpretation and communication of risk levels [8, 9]. Thus, a structured comparison of the diagnostic performance of the systems would provide insight into optimal assessment strategies [9].

Prior systematic reviews have investigated the diagnostic performance of the various ultrasound risk stratification systems for thyroid nodules [10–13]. However, traditional methods for meta-analysis cannot be used to indirectly compare multiple diagnostic techniques from separate studies. Network meta-analysis is a method to combine indirect and direct data comparing multiple treatment or testing options across studies and, in turn, to allow ranking of such options [14]. Therefore, the aim of the present study was to perform a network meta-analysis to compare the diagnostic performance of six risk stratification systems used to evaluate thyroid nodules on ultrasound for the detection of thyroid cancer.

Evidence Acquisition

Search Strategy and Study Selection

This network meta-analysis is reported in accordance with the PRISMA reporting guidelines. The research protocol was registered with the Open Science Framework.

A literature search was conducted related to the following question, which was structured using the population, intervention, comparison, and outcomes (PICO) format: In patients undergoing ultrasound for thyroid nodule evaluation (population) using any of the AACE/ACE/AME system, ACR TI-RADS, the ATA system, EU-TIRADS, Kwak TIRADS, or K-TIRADS (intervention), how did the system perform in comparison with one or more of the other systems (comparison) in terms of diagnostic performance for the detection of thyroid cancer based on a cytologic or histologic reference standard (outcome)? The PubMed, Embase, Cochrane Library, Web of Science, and Google Scholar databases were searched from inception of the databases to the date of search on August 31, 2022. A series of searches of each database was conducted using terms related to the following broad themes: thyroid neoplasms, ultrasound, biopsy, fine needle, risk assessment, risk stratification, the AACE/ACE/AME system, ACR TI-RADS, the ATA guidelines, EU-TIRADS, K-TIRADS, and Kwak-TI-RADS. The search strategy was designed with the assistance of a librarian (nonauthor) with 10 years of experience in information searches. The bibliographies of relevant studies identified by the initial searches were reviewed to identify potential additional studies. The detailed search strategy is presented in Table S1 (available in the online supplement).

Two investigators (M.A.B. and J.L., both head and neck surgeons) independently reviewed the search results to identify studies for inclusion in the analysis. Discrepancies were resolved by discussion with a third investigator (S.W.K., a head and neck surgeon). First, duplicate articles were removed. Then, the titles and abstracts of unique studies were screened to exclude articles that did not present original research (e.g., review articles and case reports) as well as to exclude original research studies that were not relevant to the study question. After this screening was completed, the full text of potentially eligible studies was retrieved and reviewed for further eligibility. On the basis of this review, additional articles were excluded if they were not relevant to the study question or if they did not provide sufficient data to calculate sensitivity and specificity. This process resulted in a final sample of studies that compared at least any two of the six ultrasound risk stratification systems and that provided sufficient data to determine the diagnostic performance for detection of thyroid cancer, with cytologic or histologic evaluation used as the reference standard.

Data Extraction

The two previously noted investigators (M.A.B. and J.L.) extracted data from included studies, consulting the third investigator (S.W.K.) to resolve discrepancies. Recorded study characteristics included the country of origin, number of patients, summary metric of patient age, distribution of patient sex, number of nodules, summary metric of nodule size, ultrasound risk stratification systems compared, details of ultrasound examinations, details of individuals interpreting ultrasound examinations, and the reference standard for thyroid malignancy. These investigators also extracted at the nodule level the number of true-positive, true-negative, false-positive, and false-negative assessments for diagnosis of thyroid cancer based on the reference standard, for all stratification systems and risk category thresholds analyzed by the study. For example, if a study reported diagnostic performance data at both intermediate-suspicion and high-suspicion thresholds for a given system, then diagnostic performance data were extracted for both of these thresholds for the system.

ACR TI-RADS and Kwak TIRADS are score-based systems whereby the risk category is determined by assigning points for the presence of certain ultrasound findings; the other four sys-

HIGHLIGHTS

Key Finding

- *Network meta-analysis of six risk stratification systems for evaluating thyroid nodules on ultrasound found both sensitivity and specificity to be highest for ACR TI-RADS (evaluated at a threshold category of TR5, denoting highly suspicious), followed by K-TIRADS (evaluated at a threshold category of K-TIRADS 5, indicating highly suspicious).*

Importance

- *ACR TI-RADS may provide optimal risk stratification, balancing cancer detection with limiting the frequency of FNA of benign nodules.*

tems are pattern-based systems whereby the risk category is determined on the basis of the characteristic overall patterns that reflect combinations of ultrasound features. The AACE/ACE/AME system classifies thyroid nodules as low risk (class 1), intermediate risk (class 2), or high risk (class 3) for malignancy [15]. ACR TI-RADS classifies thyroid nodules as benign (TR1, 0 points) or as not suspicious (TR2, 2 points), mildly suspicious (TR3, 3 points), moderately suspicious (TR4, 4–6 points), or highly suspicious (TR5, 37 points) for malignancy [6]. The ATA guidelines classify thyroid nodules as benign; as very low, low, intermediate, or high suspicion for malignancy; or as not specified [1]. EU-TIRADS classifies thyroid nodules as benign (EU-TIRADS 2) or as low (EU-TIRADS 3), intermediate (EU-TIRADS 4), or high risk for malignancy (EU-TIRADS 5) [4]. K-TIRADS classifies thyroid nodules as benign (K-TIRADS 2) or as low (K-TIRADS 3), intermediate (K-TIRADS 4), or high suspicion (K-TIRADS 5) for malignancy [5]. Kwak TIRADS classifies thyroid nodules as benign (category 2), probably benign (category 3, no suspicious ultrasound features), low suspicion for malignancy (category 4a, one suspicious ultrasound feature), intermediate suspicion for malignancy (category 4b, two suspicious ultrasound features), moderate concern but not classic for malignancy (category 4c, three or four suspicious ultrasound features), or highly suggestive of malignancy (category 5, five suspicious ultrasound features) [7]. Table 1 summarizes each system's risk categories and size threshold for FNA for each category.

Quality Assessment

The two previously noted investigators (M.A.B. and J.L.) performed a quality assessment of included studies using the Newcastle-Ottawa Scale for case control studies [16], consulting the third investigator (S.W.K.) to resolve discrepancies. Using this scale, each study was assessed for four items related to selection (adequacy of case definition, representativeness of the cases, selection of controls, and definition of controls), one item related to comparability (comparability of cases and controls on the basis of the design or analysis), and three items related to exposure (ascertainment of exposure, same method of ascertainment for cases and controls, and nonresponse rate). Up to two stars could be awarded for the single comparability item, and up to one star could be awarded for the other eight items, allowing a maximum of nine possible stars per study. Overall study quality was categorized as poor when 4 or fewer stars were awarded, fair when 5 or 6 stars were awarded, and good when 7 or more stars were awarded.

TABLE 1: Summary of Risk Stratification Systems for Assessing Thyroid Nodules on Ultrasound Evaluated in Network Meta-Analysis

Risk Stratification System	Approach	Risk Categories (Size Threshold for Biopsy by Category)
AACE/ACE/AME	Pattern based	Class 1, low risk (no biopsy) Class 2, intermediate risk (≥ 20 mm) Class 3, high risk (≥ 10 mm)
ACR TI-RADS	Score based	TR1, benign (no biopsy) TR2, not suspicious (no biopsy) TR3, mildly suspicious (≥ 25 mm; follow-up ultrasound if ≥ 15 mm) TR4, moderately suspicious (≥ 15 mm; follow-up ultrasound if ≥ 10 mm) TR5, highly suspicious (≥ 10 mm; follow-up ultrasound if ≥ 0.5 mm)
ATA	Pattern based	Not specified Benign (no biopsy) Very low suspicion (≥ 20 mm) Low suspicion (≥ 15 mm) Intermediate suspicion (≥ 10 mm) High suspicion (> 10 mm)
EU-TIRADS	Pattern based	EU-TIRADS 2, benign (no biopsy) EU-TIRADS 3, low risk (≥ 20 mm) EU-TIRADS 4, intermediate risk (≥ 15 mm) EU-TIRADS 5, high risk (> 10 mm; FNA or active surveillance if ≤ 10 mm)
K-TIRADS	Pattern based	K-TIRADS 2, benign (≥ 20 mm if spongiform) K-TIRADS 3, low suspicion (≥ 15 mm) K-TIRADS 4, intermediate suspicion (≥ 10 mm) K-TIRADS 5, high suspicion (≥ 10 mm, 0.5 mm in select cases)
Kwak TIRADS	Score based	2, Benign (no biopsy) 3, Probably benign (no biopsy) 4a, Low suspicion for malignancy (≥ 25 mm) 4b, Intermediate suspicion for malignancy (≥ 15 mm) 4c, Moderate concern but not classic for malignancy (≥ 10 mm) 5, Highly suggestive of malignancy (≥ 10 mm)

Note—AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; ATA = American Thyroid Association risk stratification system; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; FNA = fine-needle aspiration; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

Statistical Analysis

The sensitivity, specificity, and accuracy of each of the six stratification systems were determined for the systems' respective risk category thresholds, by use of the number of true-positive, false-positive, true-negative, and false-negative assessments extracted from the individual studies. I^2 was computed as a measure of heterogeneity among studies in terms of sensitivity, specificity, and accuracy for each system and was considered to be substantial when it was at least 75%. A random-effects network meta-analysis was then performed within the frequentist framework to compare the performance metrics among the six systems. For each system, the network meta-analysis used the risk category with the highest diagnostic accuracy, and the sensitivity and specificity at that category were used for the purpose of comparison with the other systems, on the basis of direct (i.e., comparisons within the same study) and indirect (i.e., comparisons using data from different studies) effect sizes [17]. The system evaluated by the largest number of studies was identified, and ORs and 95% CIs were used to express the sensitivity, specificity, and accuracy with respect to that system for the remaining five systems. The loop-specific approach was applied to identify heterogeneity between direct and indirect effect sizes in the network meta-analysis model [18], with design-by-treatment interaction models used to identify global differences and with node-splitting models used to identify local differences between pairs of systems. In design-by-treatment interaction models, the Wald test was used to identify differences between direct and indirect effect sizes by testing the linearity of regression coefficients for the entire model after calculating regression coefficients for the model for each of the six stratification systems individually. In the node-splitting models, heterogeneity between direct and indirect effect sizes was considered to be present when the 95% CI of the difference between these effect sizes excluded zero. If direct and indirect effect sizes were found to be significantly different, then subanalysis of the diagnostic performance of the six systems was conducted, selecting a parameter showing variation among studies. Subsequently, the surface under the cumulative ranking curve (SUCRA) was used to rank the six systems in terms of sensitivity, specificity, and accuracy [19]. SUCRA, which ranges from 0% to 100%, provides a single value to represent the hierarchic ranking of an option within a network meta-analysis and is dependent on the systems being compared; it therefore has no extrinsic meaning external to the given model. Finally, comparison-adjusted funnel plots were constructed to assess potential publication bias for each of sensitivity, specificity, and accuracy [20]. Asymmetry of the funnel plots, as evidence of publication bias, was assessed both visually and through linear regression analysis. The network meta-analysis was conducted using the R package netmeta (version 3.5.0).

Evidence Synthesis

Study Selection

The initial search of multiple databases identified 2618 articles, including 475 articles that were unique after duplicate records were removed. After unique articles were screened on the basis of titles and abstracts to determine their relevance to the study question, an attempt was made to retrieve the full text of 63 potentially eligible articles, 57 of which were successfully retrieved. Of these, six articles were excluded due to lack of relevance to the study question, and 12 were excluded due to lack of sufficient data to determine diagnostic performance. These exclusions resulted in inclusion of 39 studies in the final analysis [21–59]. Figure 1 shows the flow of study selection. These studies included a total of 46,661 patients with a total of 51,848 evaluated thyroid nodules. The characteristics of the included studies are summarized in Table S2 (available in the online supplement). The ultrasound examinations in the 39 studies were interpreted by practicing radiologists in 22, by a combination of practicing and in-training radiologists in five, by endocrinologists in four, and by physicians of unstated medical specialties in six; details about the individuals who interpreted the examinations were unclear in the remaining two studies.

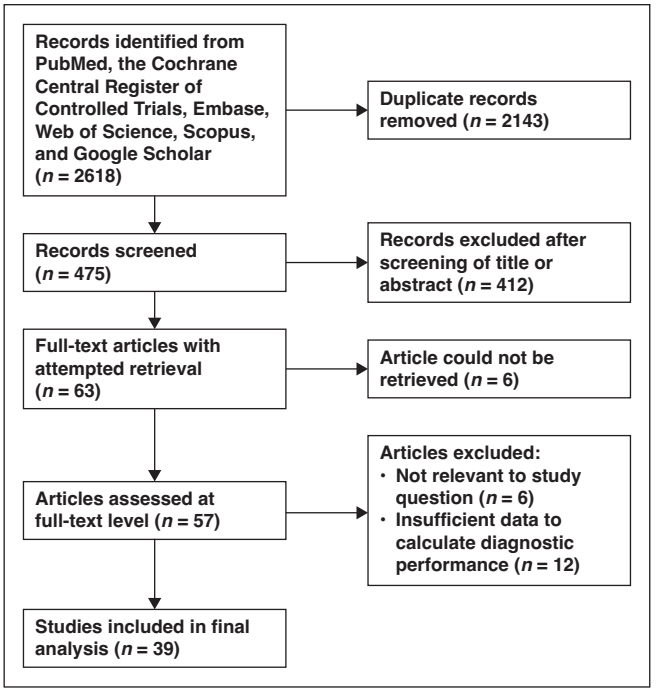


Fig. 1—Flow diagram shows study selection process. Count of records from initial search includes any articles retrieved by review of reference lists of potentially relevant articles.

Quality Assessment

Table 2 summarizes the results of the quality assessment of included studies. According to the Newcastle-Ottawa Scale, no studies had poor quality, 17 studies had fair quality, and 22 studies had good quality. A total of 23–39 studies received a star for the four questions in the selection category, with 23 studies receiving a star for all four of these questions. Nine studies received two stars for the single question in the comparability category; the remaining 30 studies received no stars for this question. A total of 30–39 studies received a star for the three questions in the exposure category, with 30 studies receiving a star for all three of these questions.

Meta-Analysis of Individual Risk Stratification Systems

Table 3 shows the diagnostic performance of the six stratification systems as determined at each system's risk category thresholds.

TABLE 2: Results of Quality Assessment of Studies Using Newcastle-Ottawa Scale for Case-Control Studies

	Selection				Comp	Exposure				
First Author [Reference]	1	2	3	4	5	6	7	8	Total	Quality
Ahmadi [21]	1	0	0	1	2	1	1	1	7	Good
Xu [22]	1	1	1	1	0	1	1	0	6	Fair
Gao [23]	1	1	1	1	0	1	1	1	7	Good
Barbosa [24]	1	1	1	1	0	1	1	1	7	Good
Chen [25]	1	0	0	1	2	1	1	1	7	Good
Chen [26]	1	1	1	1	0	1	1	0	6	Fair
Chng [27]	1	1	1	1	0	1	1	0	6	Fair
Ha [28]	1	1	1	1	0	1	1	1	7	Good
Ha [29]	1	0	0	1	2	1	1	1	7	Good
Ha [30]	1	0	1	1	0	1	1	1	6	Fair
Ha [31]	1	0	1	1	0	1	1	1	6	Fair
Ha [32]	1	0	1	1	0	1	1	1	6	Fair
Hekimsoy [33]	1	0	1	1	0	1	1	1	6	Fair
Hong [34]	1	0	1	1	0	1	1	1	6	Fair
Huang [35]	1	1	1	1	0	1	1	1	7	Good
Huh [36]	1	0	0	1	2	1	1	1	7	Good
Kang [37]	1	1	1	1	0	1	1	0	6	Fair
Li [38]	1	0	0	1	2	1	1	1	6	Fair
Lin [39]	1	1	1	1	0	1	1	1	7	Good
Na [40]	1	0	0	1	2	1	1	1	7	Good
Persichetti [41]	1	1	1	1	0	1	1	0	6	Fair
Qi [42]	1	1	1	1	0	1	1	1	7	Good
Qi [43]	1	1	1	1	0	1	1	1	7	Good
Ruan [44]	1	0	0	1	2	1	1	1	7	Good
Scappaticcio [45]	1	1	1	1	0	1	1	1	7	Good
Seifert [46]	1	1	1	1	0	1	1	1	7	Good
Shen [47]	1	1	1	1	0	1	1	1	7	Good
Shi [48]	1	1	1	1	0	1	1	1	7	Good
Wu [49]	1	1	1	1	0	1	1	1	7	Good
Xiang [50]	1	1	1	1	0	1	1	0	6	Fair
Yang [51]	1	0	0	1	2	1	1	1	7	Good
Yoo [52]	1	0	1	1	0	1	1	1	6	Fair
Yoon [53]	1	1	1	1	0	1	1	0	6	Fair
Yoon [54]	1	0	1	1	0	1	1	1	6	Fair
Yoon [55]	1	1	1	1	0	1	1	0	6	Fair
Zhang [56]	1	0	0	1	2	1	1	1	7	Good
Zhang [57]	1	1	1	1	0	1	1	1	7	Good
Zhang [58]	1	1	1	1	0	1	1	0	6	Fair
Zhu [59]	1	1	1	1	0	1	1	1	7	Good

Note—The Newcastle-Ottawa scale for case-control studies is described in full by Wells et al. [16]. Except where otherwise indicated, values denote number of stars awarded to a study for each item. In column headings, Comp = comparability, 1 = adequacy of case definition, 2 = representativeness of cases, 3 = selection of controls, 4 = definition of controls, 5 = comparability of cases and controls based on the design or analysis, 6 = ascertainment of exposure, 7 = same method of ascertainment used for cases and controls, 8 = nonresponse rate.

TABLE 3: Results of Meta-Analysis of Diagnostic Performance of Individual Risk Stratification Systems

System and Category	Sensitivity		Specificity		Accuracy	
	% (95% CI)	I^2	% (95% CI)	I^2	% (95% CI)	I^2
ATA						
Low suspicion	97 (95–98)	97.5	26 (19–35)	99.5	54 (47–61)	99.4
Intermediate suspicion	87 (80–91)	98.2	64 (56–71)	99.3	72 (67–76)	99.0
High suspicion	71 (60–79)	98.9	86 (78–91)	99.5	79 (74–84)	99.1
AACE/ACE/AME						
Class 1, low risk	99 (96–99)	90.3	0 (0–1)	97.9	18 (15–21)	92.0
Class 2, intermediate risk	98 (97–99)	77.6	6 (3–11)	98.4	24 (19–29)	96.9
Class 3, high risk	64 (45–80)	98.6	85 (76–91)	99.1	80 (76–83)	95.1
ACR TI-RADS						
TR3, mildly suspicious	98 (97–99)	96.8	23 (17–30)	99.5	52 (47–56)	99.1
TR4, moderately suspicious	93 (89–95)	98.0	54 (48–60)	99.1	68 (64–71)	98.5
TR5, highly suspicious	65 (55–74)	99.2	89 (86–92)	98.7	81 (78–83)	97.2
EU-TIRADS						
EU-TIRADS 3, low risk	99 (98–99)	97.6	3 (1–7)	99.3	32 (25–39)	99.3
EU-TIRADS 4, intermediate risk	93 (89–95)	95.3	49 (41–58)	99.3	62 (57–68)	98.9
EU-TIRADS 5, high risk	71 (61–80)	98.3	82 (75–87)	99.3	78 (74–81)	97.0
K-TIRADS						
K-TIRADS 3, low suspicion	99 (99–99)	82.7	7 (3–15)	99.5	43 (33–54)	99.7
K-TIRADS 4, intermediate suspicion	92 (88–95)	98.2	63 (58–69)	99.0	73 (69–77)	98.9
K-TIRADS 5, high suspicion	65 (55–74)	98.9	90 (86–93)	98.8	83 (80–85)	97.8
Kwak TIRADS						
4a, Low suspicion for malignancy	99 (97–99)	96.8	31 (24–40)	98.7	61 (54–68)	98.9
4b, Intermediate suspicion for malignancy	96 (93–98)	95.3	56 (50–63)	97.6	74 (70–78)	96.9
4c, Moderate concern but not classic for malignancy	77 (66–86)	98.7	83 (79–87)	97.3	81 (76–85)	98.4
5, Highly suggestive of malignancy	14 (10–18)	95.0	99 (98–99)	87.8	64 (55–71)	99.1

Note—ATA = American Thyroid Association risk stratification system; AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

Five studies evaluated the AACE/ACE/AME system [28, 30, 39–41]. Sensitivity ranged from 64% (class 3, high risk) to 99% (class 1, low risk). Specificity ranged from 0% (class 1, low risk) to 85% (class 3, high risk). Accuracy ranged from 18% (class 1, low risk) to 80% (class 3, high risk).

Thirty-two studies evaluated ACR TI-RADS [21–26, 28–31, 33, 35–37, 39, 40, 42–49, 51, 52, 54–59]. Sensitivity ranged from 65% (TR5, highly suspicious) to 98% (TR3, mildly suspicious). Specificity ranged from 23% (TR3, mildly suspicious) to 89% (TR5, highly suspicious). Accuracy ranged from 52% (TR3, mildly suspicious) to 81% (TR5, highly suspicious).

Thirty studies evaluated the ATA guidelines [21, 23, 24, 26–29, 31, 32, 34–36, 38–41, 43–47, 49–54, 56, 57, 59]. Sensitivity ranged from 71% (high suspicion) to 97% (low suspicion). Specificity ranged from 26% (low suspicion) to 86% (high suspicion). Accuracy ranged from 54% (low suspicion) to 79% (high suspicion).

Fourteen studies evaluated EU-TIRADS [22, 30, 33, 36, 39, 40, 42, 45–48, 52, 54, 55]. Sensitivity ranged from 71% (EU-TIRADS 5, high risk) to 99% (EU-TIRADS 3, low risk). Specificity ranged from 3% (EU-TIRADS 3, low risk) to 82% (EU-TIRADS 5, high risk). Accuracy ranged from 32% (EU-TIRADS 3, low risk) to 78% (EU-TIRADS 5, high risk).

Twenty-three studies evaluated K-TIRADS [22, 25, 26, 28–32, 34, 37, 39, 40, 42, 45, 46, 50–52, 54–57, 59]. Sensitivity ranged from 65% (K-TIRADS 5, high suspicion) to 99% (K-TIRADS 3, low suspicion). Specificity ranged from 7% (K-TIRADS 3, low suspicion) to 90% (K-TIRADS 5, high suspicion). Accuracy ranged from 43% (K-TIRADS 3, low suspicion) to 83% (K-TIRADS 5, high suspicion).

Fourteen studies evaluated Kwak TIRADS [22, 23, 27, 36, 38, 42, 43, 46–48, 53, 54, 56, 57]. Sensitivity ranged from 14% (category 5, highly suggestive of malignancy) to 99% (category 4a, low suspicion). Specificity ranged from 31% (category 4a, low suspicion)

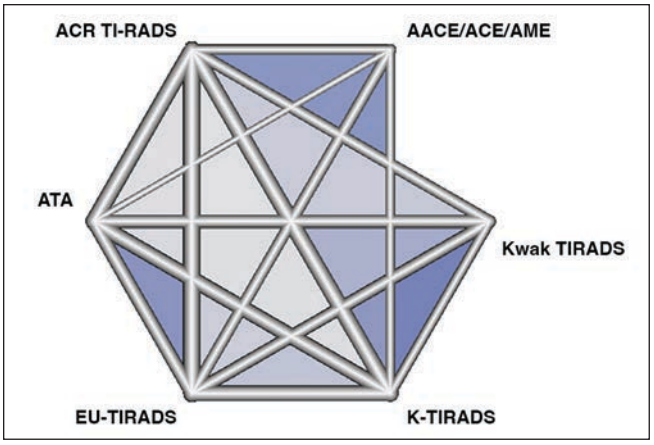


Fig. 2—Schematic of network meta-analysis of 39 studies of risk classification systems for thyroid nodules on ultrasound. Direct comparisons within individual studies are indicated by lines connecting pairs of systems. Number of studies involved in each pairwise comparison is indicated by width of lines. ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ATA = American Thyroid Association risk stratification system; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7]; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by Korean Thyroid Association and Korean Society of Thyroid Radiology.

to 99% (category 5, highly suggestive of malignancy). Accuracy ranged from 61% (category 4a, low suspicion) to 81% (category 4C, moderate concern but not classic). Category 4C had a sensitivity of 77% and a specificity of 83%.

Heterogeneity was substantial for sensitivity, specificity, and accuracy for all assessed risk category thresholds for all systems (all $I^2 > 75\%$).

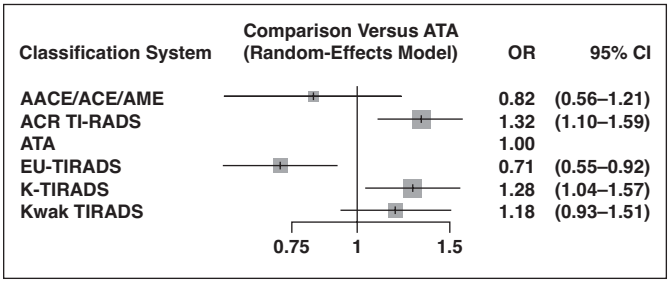
Network Meta-Analysis Based on Selected Risk Category Thresholds for Each System

The network meta-analysis [21–59] was performed using threshold categories (based on highest accuracy) of class 3 (high risk) for the AACE/ACE/AME system, TR5 (highly suspicious) for ACR TI-RADS, EU-TIRADS 5 (high risk) for EU-TIRADS, 4c (moderate concern but not classic for malignancy) for Kwak TIRADS, K-TIRADS 5 (high suspicion) for K-TIRADS, and high suspicion for the ATA system. Figure 2 depicts the direct comparisons within the network meta-analysis. The ATA system was selected as the reference system for the network meta-analysis because this system was examined by the largest number of included studies. Figure 3 graphically depicts the results of the network meta-analysis in terms of the OR for comparison of the sensitivity, specificity, and accuracy of the ATA system with those of each of the five other systems.

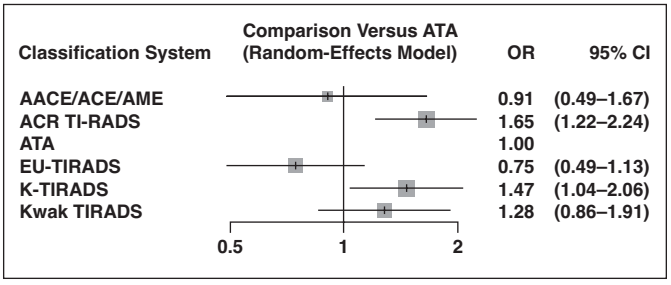
Sensitivity was significantly higher than the ATA system for ACR TI-RADS (OR = 1.32 [95% CI, 1.10–1.59]), K-TIRADS (OR = 1.28 [95% CI, 1.04–1.57]), and higher (although not significantly) for Kwak TIRADS (OR = 1.18 [95% CI, 0.93–1.51]). Compared with the sensitivity of the ATA system, sensitivity was significantly lower for EU-TIRADS (OR = 0.71 [95% CI, 0.55–0.92]) and lower (although not significantly) for the AACE/ACE/AME system (OR = 0.82 [95% CI, 0.56–1.21]).

Compared with the specificity of the ATA system, specificity was significantly higher for ACR TI-RADS (OR = 1.65 [95% CI, 1.22–2.24]) and K-TIRADS (OR = 1.47 [95% CI, 1.04–2.06]) and was higher (although not significantly) for Kwak TIRADS (OR = 1.28 [95% CI, 0.86–1.91]). Compared with the specificity of the ATA system, specificity was lower (although not significantly) for the AACE/ACE/AME system (OR = 0.91 [95% CI, 0.49–1.67]) and EU-TIRADS (OR = 0.75 [95% CI, 0.49–1.13]).

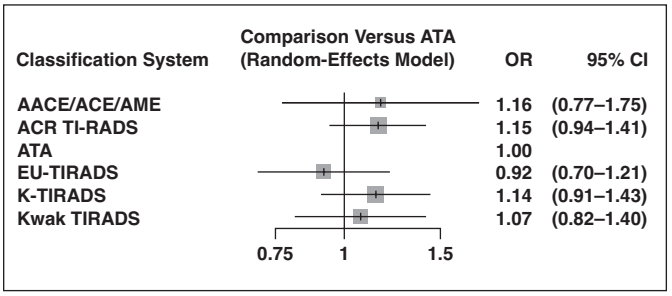
Compared with the accuracy of the ATA system, accuracy was higher (although not significantly) for ACR TI-RADS (OR = 1.15 [95% CI, 0.94–1.41]), K-TIRADS (OR = 1.14 [95% CI, 0.91–1.43]), the



A



B



C

Fig. 3—Results of network meta-analysis.
A–**C,** Charts show summary of sensitivity (**A**), specificity (**B**), and accuracy (**C**). Tick marks indicate ORs, gray boxes around tick marks are proportional to precision of estimates, and horizontal lines denote 95% CIs. American Thyroid Association (ATA) risk stratification system serves as reference for ORs. AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by Korean Thyroid Association and Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

AACE/ACE/AME system (OR = 1.16 [95% CI, 0.7775]), and Kwak TIRADS (OR = 1.07 [95% CI, 0.82–1.40]). Accuracy was lower (although not significantly) for EU-TIRADS (OR = 0.92 [95% CI, 0.70–1.21]) than for the ATA system.

According to the network meta-analysis, global differences between direct and indirect effect sizes were not significant in terms of sensitivity (p = .32), specificity (p = .36), or accuracy (p = .12). However, local direct and indirect effect sizes were significantly different for the comparison of specificity between K-TRADS and the AACE/ACE/AME system (direct: OR = 0.30 [95% CI, 0.13–0.66]; indirect: OR = 1.97 [95% CI, 0.73–5.34]; p = .004), the comparison of specificity between K-TIRADS and the ATA system (direct: OR = 1.10 [95% CI, 0.74–1.63]; indirect: OR = 3.44 [95% CI, 1.75–6.80]; p = .004), the comparison of accuracy between the AACE/ACE/AME system and ACR TI-RADS (direct: OR = 0.68 [95% CI, 0.41–1.15]; indirect: OR = 1.87 [95% CI, 0.97–3.62]; p = .02), and the comparison of accuracy between the AACE/ACE/AME system and K-TIRADS (direct: OR = 0.69 [95% CI, 0.41–1.16]; indirect: OR = 1.99 [95% CI, 1.00–3.97]; p = .02).

Subgroup Analysis

Subgroup analysis was conducted given the discrepancies between direct and indirect effect sizes. Nodule size was selected as the basis for this subanalysis given variation in minimum nodule size across studies. A total of 12 studies included only nodules measuring 1 cm or larger [21, 27, 28, 30, 36, 40, 43, 46, 48, 53–55]. The remaining 27 studies included nodules measuring less than 1 cm and thus reported results for nodules of all sizes [22–26, 29, 31–35, 37–39, 41, 42, 44, 45, 47, 49–52, 56–59]; of these, eight studies reported additional results for the subset of nodules measuring 1 cm or larger [23, 29, 31, 32, 37, 50, 56, 59], providing a total of 20 studies that reported results for nodules measuring 1 cm or larger. Table 4 shows the results of the subgroup analysis, stratifying comparisons based on the 27 studies that reported results for nodules of all sizes [22–26, 29, 31–35, 37–39, 41, 42, 44, 45, 47, 49–52, 56–59] and the 20 studies that reported results for nodules 1 cm or larger [21, 23, 27–32, 36, 37, 40, 43, 46, 48, 50, 53–56, 59]. In subgroup analysis of studies that reported results for nodules of all sizes, the five other systems showed greater sensitivity than the ATA system, but none of these differences were statistically significant. In comparison, in subgroup analysis of studies that reported results for nodules 1 cm or larger, ACR TI-RADS (OR = 1.75 [95% CI, 1.45–2.12]) and K-TIRADS (OR = 1.70 [95% CI, 1.37–2.10]) showed significantly greater sensitivity than the ATA system. In subgroup analysis of studies that reported results for nodules of all sizes, the AACE/ACE/AME system (OR = 4.16 [95% CI, 1.33–13.03]) and ACR TI-RADS (OR = 2.14 [95% CI, 1.41–3.26]) showed significantly higher specificity than the ATA system. In comparison, in a subgroup analysis of studies that reported results for nodules 1 cm or larger, ACR TI-RADS (OR = 1.44 [95% CI, 1.18–1.77]) and K-TIRADS (OR = 1.30 [95% CI, 1.04–1.62]) showed significantly higher specificity than the ATA system, and the AACE/ACE/AME system (OR = 0.42 [95% CI, 0.29–0.60]) and EU-TIRADS (OR = 0.52 [95% CI, 0.40–0.66]) showed significantly lower specificity than the ATA system. In subgroup analysis of studies that reported results for nodules of all sizes, the AACE/ACE/AME system (OR = 3.27 [95% CI, 1.52–7.06]) and ACR TI-RADS (OR = 1.42 [95% CI, 1.06–1.90]) showed significantly higher accuracy than the ATA system. In comparison, in subgroup analysis of studies that reported results

TABLE 4: Subanalysis of Risk Stratification System Based on Studies Reporting Results for Nodules of All Sizes and Studies Reporting Results for Nodules 1 cm or Larger

Risk Stratification System	Sensitivity	Specificity	Accuracy
AACE/ACE/AME			
All sizes	1.75 (0.64–4.71)	4.16 (1.33–13.03)	3.27 (1.52–7.06)
≥ 1 cm	0.86 (0.63–1.18)	0.42 (0.29–0.60)	0.61 (0.47–0.78)
ACR TI-RADS			
All sizes	1.29 (0.94–1.76)	2.14 (1.41–3.26)	1.42 (1.06–1.90)
≥ 1 cm	1.75 (1.45–2.12)	1.44 (1.18–1.77)	1.00 (0.87–1.15)
EU-TIRADS			
All sizes	0.81 (0.51–1.30)	1.53 (0.81–2.92)	1.51 (0.98–2.34)
≥ 1 cm	0.85 (0.67–1.08)	0.52 (0.40–0.66)	0.63 (0.53–0.75)
K-TIRADS			
All sizes	1.21 (0.87–1.69)	1.54 (0.99–2.41)	1.24 (0.91–1.69)
≥ 1 cm	1.70 (1.37–2.10)	1.30 (1.04–1.62)	0.97 (0.83–1.13)
Kwak TIRADS			
All sizes	1.24 (0.79–1.95)	1.55 (0.84–2.85)	1.21 (0.78–1.87)
≥ 1 cm	1.24 (1.00–1.55)	1.11 (0.88–1.40)	0.97 (0.82–1.14)

Note—Data are expressed OR with 95% CI in parentheses, as calculated with respect to performance of the American Thyroid Association (ATA) risk stratification system. AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

TABLE 5: Rankings of Sensitivity, Specificity, and Accuracy of Risk Classification Systems Based on Network Meta-Analysis

System	Sensitivity		Specificity		Accuracy	
	SUCRA	Rank	SUCRA	Rank	SUCRA	Rank
AACE/ACE/AME	20	5	27	5	66	3
ACR TI-RADS	89	1	93	1	72	1
ATA	39	4	33	4	30	5
EU-TIRADS	5	6	8	6	14	6
K-TIRADS	81	2	78	2	68	2
Kwak TIRADS	67	3	62	3	50	4

Note—Surface under the cumulative ranking curve (SUCRA) values are percentages. AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; ATA = American Thyroid Association risk stratification system; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

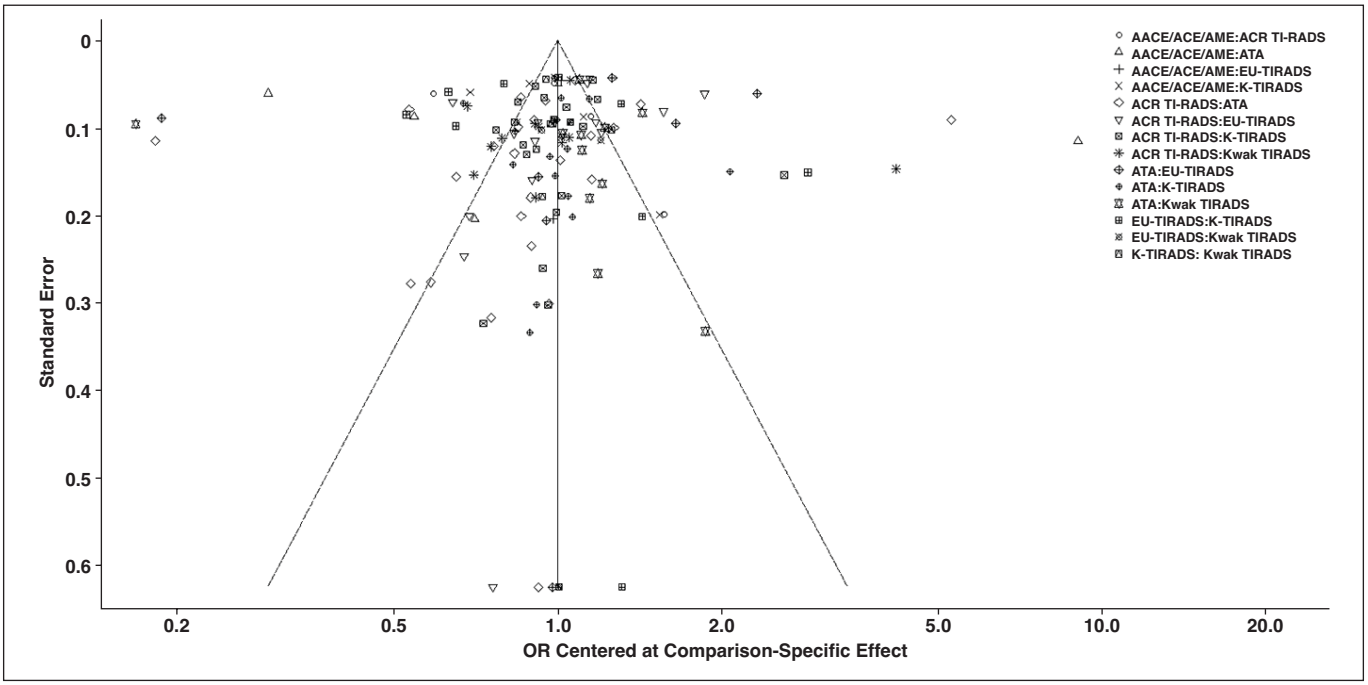
for nodules 1 cm or larger, the AACE/ACE/AME system (OR = 0.61 [95% CI, 0.47–0.78]) and EU-TIRADS (OR = 0.63 [95% CI, 0.53–0.75]) showed significantly lower specificity than the ATA system.

System Rankings

The SUCRA values for the six risk stratification systems are presented in Table 5. SUCRA was highest for ACR TI-RADS for sensitivity (89%), specificity (93%), and accuracy (72%). SUCRA was second highest for K-TIRADS for sensitivity (81%), specificity (78%), and accuracy (68%). SUCRA was lowest for EU-TIRADS for sensitivity (5%), specificity (8%), and accuracy (14%).

Publication Bias

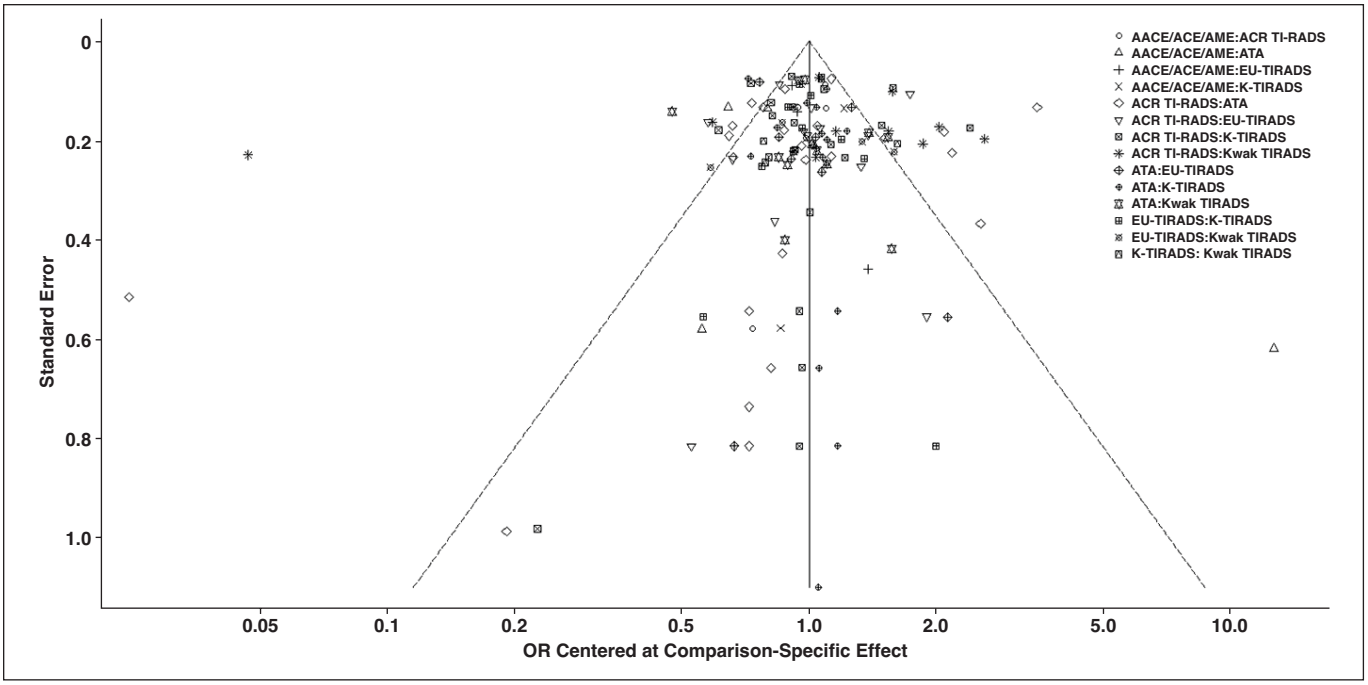
Figure 4 shows the comparison-adjusted funnel plots of sensitivity, specificity, and accuracy for the six classification systems. The funnel plots appeared symmetric, without visual indication



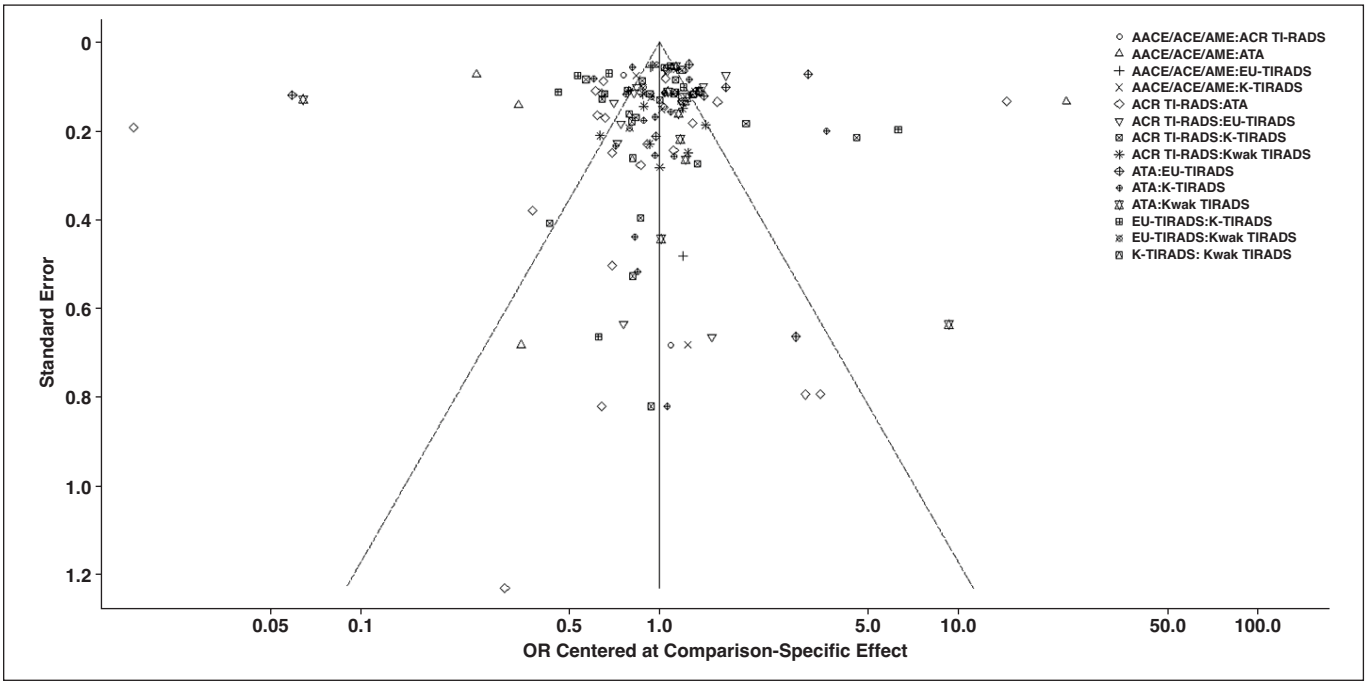
A

Fig. 4—Comparison-adjusted funnel plots of six classification systems.
A–**C,** Funnel plots show assessments for publication bias with respect to accuracy (**A**), sensitivity (**B**), and specificity (**C**), with no evidence of publication bias provided for any of three measures. AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; ATA = American Thyroid Association risk stratification systems; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by Korean Thyroid Association and Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

(Fig. 4 continues on next page)



B



C

Fig. 4 (continued)—Comparison-adjusted funnel plots of six classification systems.
A–**C,** Funnel plots show assessments for publication bias with respect to accuracy (**A**), sensitivity (**B**), and specificity (**C**), with no evidence of publication bias provided for any of three measures. AACE/ACE/AME = American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi system; ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System; ATA = American Thyroid Association risk stratification systems; EU-TIRADS = European Thyroid Association Thyroid Imaging Reporting and Data System; K-TIRADS = Korean Thyroid Imaging Reporting and Data System endorsed by Korean Thyroid Association and Korean Society of Thyroid Radiology; Kwak TIRADS = Thyroid Imaging Reporting and Data System developed by Kwak et al. [7].

of publication bias. Linear regression analysis also indicated an absence of asymmetry for sensitivity ($p = .92$), specificity ($p = .96$), and accuracy ($p = .81$); thus, they also did not provide evidence of publication bias.

Discussion

This network meta-analysis compared diagnostic performance among six ultrasound-based risk stratification systems used for diagnosing thyroid malignancy. Meta-analysis of individual systems identified the threshold categories having the highest accuracy as class 3 (high risk) for the AACE/ACE/AME system, TR5 (highly suspicious) for ACR TI-RADS, high suspicion for the ATA system, EU-TIRADS 5 (high risk) for EU-TIRADS, category 4c (moderate concern but not classic for malignancy) for Kwak TIRADS, and K-TIRADS 5 (high suspicion) for K-TIRADS. At these category thresholds, the risk stratification systems had sensitivity of 64–77% and specificity of 82–90%. The network meta-analysis found both highest sensitivity and highest specificity, based on SUCRA values determined for these category thresholds, for ACR TI-RADS, followed by K-TIRADS.

Thyroid nodules undergo FNA to provide a definitive diagnosis of malignancy. However, obtaining a definitive diagnosis must be balanced with limiting the frequency of FNA of benign nodules. Accordingly, organizations have developed and revised ultrasound risk stratification systems. The classifications provide practical scoring systems that aim to select nodules for biopsy on the basis of nodule size and risk of malignancy, facilitating communication between radiologists and referring clinicians. The six such systems explored in the present investigation (the AACE/ACE/AME system, ACR TI-RADS, the ATA guideline, EU-TIRADS, Kwak TIRADS, and K-TIRADS) differ in terms of the type and number of risk categories that may be assigned to detected nodules as well as in terms of thyroid cancer risk estimates. Furthermore, the six systems define risk categories using varying ultrasound findings. Thus, exploration of their relative diagnostic performance is helpful to guide clinical implementation.

ACR TI-RADS and Kwak TIRADS are score-based systems, whereas the other four systems are pattern based. The pattern-based systems are simple and easy to apply clinically, but they provide less precise estimates of malignancy risk [55]. For example, pattern-based systems do not consider a solid component as a risk factor distinct from other suspicious ultrasound findings. Indeed, prior work has shown higher overall accuracy and AUC of score-based systems [54]. Kwak TIRADS, which is the simpler of the two score-based systems, determines the risk of malignancy on the basis of the number of suspicious ultrasound features, weighing all features equally. In comparison, ACR TI-RADS entails initial assignment of a varying number of points in multiple distinct categories before calculating the sum of these points across categories in a manner that more strongly weighs certain findings. In addition, ACR TI-RADS considers commonly encountered thyroid nodule characteristics (e.g., regular shape and margins, mild hypoechogenicity, and mixed composition) to be mildly suspicious, such that nodules with these characteristics are generally assigned risk scores lower than those in other systems. Such factors may account for the high rankings of sensitivity and specificity of ACR TI-RADS in the present network meta-analysis. An additional consideration is that the size thresholds for selecting mildly and moderately suspicious nodules to undergo FNA by ACR TI-RADS (2.5 and 1.5 cm, respectively) are larger than the size thresholds for the other risk stratification systems [43]. Thus, some malignant nodules may be selected to undergo FNA by the other systems but not by ACR TI-RADS, despite the higher performance of the ACR TI-RADS risk categories themselves. On the other hand, ACR TI-RADS is unique in recommending that follow-up ultrasound be performed for nodules smaller than the size cutoffs in the mildly, moderately, and highly suspicious categories.

This network meta-analysis had limitations. First, diagnostic performance may have been affected by uncontrolled variables such as quality of ultrasound equipment, scanning technique, interpreter experience, and availability of clinical information to those interpreting the examinations. However, the large-scale nature of the analysis reduces the impact of selection bias in individual studies [18]. Integration of studies with varying populations, designs, and reference standards, along with the synthesis of direct and indirect comparisons among studies performed under a range of conditions, also introduces uncertainty into the observations. For example, discrepant results were observed between direct and indirect effect sizes among studies. This discrepancy was further explored by subgroup analysis, which identified that, among studies, variation in minimum nodule size was a factor contributing to this discrepancy. In addition, only studies with a cytologic or histologic reference standard were included; results may have differed if patients with only clinical and imaging follow-up had been included. Also, the systems were compared in terms of the diagnostic performance of the risk category thresholds; actual recommendations for biopsy based on combinations of risk category and nodule size were not assessed. Another limitation is that the extent of overlap in patient samples among included studies, as well as the potential effect of such overlap, is unclear. Finally, the systems were compared on the basis of a single category threshold per system; these individual selected thresholds per system do not fully reflect real-world experience in applying the systems using a range of risk categories.

Conclusion

In this network meta-analysis of six risk stratification systems for evaluating thyroid nodules on ultrasound, sensitivity and specificity were highest for ACR TI-RADS (evaluated at a threshold category of TR5, highly suspicious), followed by K-TIRADS (evaluated using a threshold category of K-TIRADS 5, highly suspicious). This comparative evaluation of risk stratification systems for thyroid nodules can inform decisions regarding system implementation as well as aid future system updates.

Provenance and review: Not solicited; externally peer reviewed.

Peer reviewers: Irmak Durur-Subasi, Istanbul Medipol University; Luyao Shen, Stanford University School of Medicine; additional individuals who chose not to disclose their identities.

References

1. Haugen BR, Alexander EK, Bible KC, 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
2. Kitahara CM, Sosa JA. The changing incidence of thyroid cancer. *Nat Rev*

Endocrinol 2016; 12:646–653
3. Morris LG, Tuttle RM, Davies L. Changing trends in the incidence of thyroid cancer in the United States. *JAMA Otolaryngol Head Neck Surg* 2016; 142:709–711
4. 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
5. Shin JH, Baek JH, Chung J, 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
6. 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
7. Kwak JY, Han KH, Yoon JH, et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. *Radiology* 2011; 260:892–899
8. Yim Y, Na DG, Ha EJ, et al. Concordance of three international guidelines for thyroid nodules classified by ultrasonography and diagnostic performance of biopsy criteria. *Korean J Radiol* 2020; 21:108–116
9. Yoon JH, Han K, Kim EK, Moon HJ, Kwak JY. Diagnosis and management of small thyroid nodules: a comparative study with six guidelines for thyroid nodules. *Radiology* 2017; 283:560–569
10. Castellana M, Castellana C, Treglia G, et al. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. *J Clin Endocrinol Metab* 2020; 105:1659–1669
11. Li W, Wang Y, Wen J, Zhang L, Sun Y. Diagnostic performance of American College of Radiology TI-RADS: a systematic review and meta-analysis. *AJR* 2021; 216:38–47
12. Staibano P, Ham J, Chen J, Zhang H, Gupta MK. Inter-rater reliability of thyroid ultrasound risk criteria: a systematic review and meta-analysis. *Laryngoscope* 2022 Aug 30 [published online]
13. Yang R, Zou X, Zeng H, Zhao Y, Ma X. Comparison of diagnostic performance of five different ultrasound TI-RADS classification guidelines for thyroid nodules. *Front Oncol* 2020; 10:598225
14. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. *Intern Emerg Med* 2017; 12:103–111
15. Gharib H, Papini E, Garber JR, et al.; AACE/ACE/AME Task Force on Thyroid Nodules. 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; 22:622–639
16. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital website. www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed Mar 9, 2023
17. Power M, Fell G, Wright M. Principles for high-quality, high-value testing. *Evid Based Med* 2013; 18:5–10
18. Flather MD, Farkouh ME, Pogue JM, Yusuf S. Strengths and limitations of meta-analysis: larger studies may be more reliable. *Control Clin Trials* 1997; 18:568–579; discussion, 661–666
19. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* 2011; 64:163–171
20. Higgins JP, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods* 2012; 3:98–110
21. Ahmadi S, Herbst R, Oyekunle T, et al. Using the ATA and ACR TI-RADS sonographic classifications as adjunctive predictors of malignancy for indeterminate thyroid nodules. *Endocr Pract* 2019; 25:908–917
22. Xu T, Wu Y, Wu RX, et al. Validation and comparison of three newly-released thyroid imaging reporting and data systems for cancer risk determination. *Endocrine* 2019; 64:299–307
23. Gao L, Xi X, Jiang Y, et al. Comparison among TIRADS (ACR TI-RADS and KWAK- TI-RADS) and 2015 ATA guidelines in the diagnostic efficiency of thyroid nodules. *Endocrine* 2019; 64:90–96
24. Barbosa TLM, Junior COM, Graf H, et al. ACR TI-RADS and ATA US scores are helpful for the management of thyroid nodules with indeterminate cytology. *BMC Endocr Disord* 2019; 19:112
25. Chen Q, Lin M, Wu S. Validating and comparing C-TIRADS, K-TIRADS and ACR-TIRADS in stratifying the malignancy risk of thyroid nodules. *Front Endocrinol (Lausanne)* 2022; 13:899575
26. Chen X, Kutaiba N, Pearce S, Digby S, Van Gelderen D. Application of Thyroid Imaging Reporting and Data System (TIRADS) guidelines to thyroid nodules with cytopathological correlation and impact on healthcare costs. *Intern Med J* 2022; 52:1366–1373
27. Chng CL, Tan HC, Too CW, et al. Diagnostic performance of ATA, BTA and TIRADS sonographic patterns in the prediction of malignancy in histologically proven thyroid nodules. *Singapore Med J* 2018; 59:578–583
28. 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
29. 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; 28:1532–1537 [Erratum in *Thyroid* 2019; 29:159]
30. Ha EJ, Shin JH, Na DG, 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
31. Ha SM, Baek JH, Choi YJ, et al. Malignancy risk of initially benign thyroid nodules: validation with various thyroid imaging reporting and data system guidelines. *Eur Radiol* 2019; 29:133–140
32. Ha SM, Baek JH, Na DG, et al. Diagnostic performance of practice guidelines for thyroid nodules: thyroid nodule size versus biopsy rates. *Radiology* 2019; 291:92–99
33. 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
34. Hong HS, Lee JY. Diagnostic performance of ultrasound patterns by K-TI-RADS and 2015 ATA guidelines in risk stratification of thyroid nodules and follicular lesions of undetermined significance. *AJR* 2019; 213:444–450
35. Huang BL, Ebner SA, Makkar JS, et al. A multidisciplinary head-to-head comparison of American College of Radiology Thyroid Imaging and Reporting Data System and American Thyroid Association ultrasound risk stratification systems. *Oncologist* 2020; 25:398–403
36. 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
37. Kang S, Kwon SK, Choi HS, et al. Comparison of Korean vs. American Thyroid Imaging Reporting and Data System in malignancy risk assessment of indeterminate thyroid nodules. *Endocrinol Metab (Seoul)* 2021; 36:1111–1120
38. Li J, Li H, Yang Y, Zhang X, Qian L. The KWAK TI-RADS and 2015 ATA guide-

lines for medullary thyroid carcinoma: combined with cell block-assisted ultrasound-guided thyroid fine-needle aspiration. *Clin Endocrinol (Oxf)* 2020; 92:450–460

39. Lin Y, Lai S, Wang P, et al. Performance of current ultrasound-based malignancy risk stratification systems for thyroid nodules in patients with follicular neoplasms. *Eur Radiol* 2022; 32:3617–3630
40. 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; 40:474–485
41. Persichetti A, Di Stasio E, Guglielmi R, et al. Predictive value of malignancy of thyroid nodule ultrasound classification systems: a prospective study. *J Clin Endocrinol Metab* 2018; 103:1359–1368
42. Qi Q, Zhou A, Guo S, et al. Explore the diagnostic efficiency of Chinese thyroid imaging reporting and data systems by comparing with the other four systems (ACR TI-RADS, Kwak-TIRADS, KSThR-TIRADS, and EU-TIRADS): a single-center study. *Front Endocrinol (Lausanne)* 2021; 12:763897
43. Qi TY, Chen X, Liu H, et al. Comparison of thyroid nodule FNA rates recommended by ACR TI-RADS, Kwak TI-RADS and ATA guidelines. *Eur J Radiol* 2022; 148:110152
44. Ruan JL, Yang HY, Liu RB, 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
45. Scappaticcio L, Maiorino MI, Iorio S, et al. Exploring the performance of ultrasound risk stratification systems in thyroid nodules of pediatric patients. *Cancers (Basel)* 2021; 13:5304
46. Seifert P, Schenke S, Zimny M, et al. Diagnostic performance of Kwak, EU, ACR, and Korean TIRADS as well as ATA guidelines for the ultrasound risk stratification of non-autonomously functioning thyroid nodules in a region with long history of iodine deficiency: a German multicenter trial. *Cancers (Basel)* 2021; 13:4467
47. 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
48. Shi YX, Chen L, Liu YC, et al. Differences among the Thyroid Imaging Reporting and Data System proposed by Korean, the American College of Radiology and the European Thyroid Association in the diagnostic performance of thyroid nodules. *Transl Cancer Res* 2020; 9:4958–4967
49. Wu XL, Du JR, 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
50. Xiang P, Chu X, Chen G, et al. Nodules with nonspecific ultrasound pattern according to the 2015 American Thyroid Association malignancy risk stratification system: a comparison to the Thyroid Imaging Reporting and Data System (TIRADS-Na). *Medicine (Baltimore)* 2019; 98:e17657
51. Yang W, Fananapazir G, LaRoy J, Wilson M, Campbell MJ. Can the American Thyroid Association, K-Tirads, and Acr-Tirads ultrasound classification systems be used to predict malignancy in Bethesda category IV nodules? *Endocr Pract* 2020; 26:945–952
52. Yoo WS, Ahn HY, Ahn HS, et al. Malignancy rate of Bethesda category III thyroid nodules according to ultrasound risk stratification system and cytological subtype. *Medicine (Baltimore)* 2020; 99:e18780
53. 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
54. Yoon JH, Lee HS, Kim EK, Moon HJ, Park VY, Kwak JY. Pattern-based vs. score-based guidelines using ultrasound features have different strengths in risk stratification of thyroid nodules. *Eur Radiol* 2020; 30:3793–3802
55. Yoon SJ, Na DG, Gwon HY, et al. Similarities and differences between thyroid imaging reporting and data systems. *AJR* 2019; 213:[web]W76–W84
56. Zhang WB, Li JJ, Chen XY, et al. SWE combined with ACR TI-RADS categories for malignancy risk stratification of thyroid nodules with indeterminate FNA cytology. *Clin Hemorheol Microcirc* 2020; 76:381–390
57. Zhang WB, Xu W, Fu WJ, He BL, Liu H, Deng WF. Comparison of ACR TI-RADS, Kwak TI-RADS, ATA guidelines and KTA/KSThR guidelines in combination with SWE in the diagnosis of thyroid nodules. *Clin Hemorheol Microcirc* 2021; 78:163–174
58. Zhang Z, Lin N. Clinical diagnostic value of American College of Radiology thyroid imaging report and data system in different kinds of thyroid nodules. *BMC Endocr Disord* 2022; 22:145
59. Zhu H, Yang Y, Wu S, Chen K, Luo H, Huang J. Diagnostic performance of US-based FNAB criteria of the 2020 Chinese guideline for malignant thyroid nodules: comparison with the 2017 American College of Radiology guideline, the 2015 American Thyroid Association guideline, and the 2016 Korean Thyroid Association guideline. *Quant Imaging Med Surg* 2021; 11:3604–3618

(Editorial Comment starts on next page)

Editorial Comment: Meta-Analysis Supports ACR TI-RADS for Risk Stratification of Thyroid Nodules Over Other Systems

The American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) guides the reporting of and management recommendations for incidental thyroid nodules, aiding identification of clinically significant cancers while balancing their identification with the risk of overtreatment of benign nodules or indolent cancers [1]. The current meta-analysis by Kim et al. compares ACR TI-RADS with multiple other approaches, such as the American Thyroid Association (ATA) guidelines, American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi (AACE/ACE/AME) system, Korean Thyroid Imaging Reporting and Data System (K-TIRADS) endorsed by the Korean Thyroid Association and the Korean Society of Thyroid Radiology TIRADS, European Thyroid Association Thyroid Imaging Reporting and Data System (EU-TIRADS), and Thyroid Imaging Reporting and Data System developed by Kwak et al. [2] (Kwak TIRADS).

A prior meta-analysis that included ACR TI-RADS, ATA guidelines, Kwak TIRADS, K-TIRADS, and EU-TIRADS showed similar pooled sensitivity, specificity, and diagnostic accuracy for all of these systems, with ACR TI-RADS performing the best in terms of the relative diagnostic OR [3]. In the current meta-analysis of 39 studies with 49,661 patients, Kim et al. further substantiate these results by comparing the previously noted six systems. Among the six systems, the authors found that ACR TI-RADS had the highest diagnostic performance (sensitivity and specificity) for the most suspicious category in each system.

The internal structures of these systems have differences. For example, the ATA guidelines are based on a pattern recognition approach, whereas ACR TI-RADS is a point-based system that adds the points assigned for suspicious features to form a final score [1]. Although ACR TI-RADS can be cumbersome to use compared with the ATA guidelines, the structured point system helps achieve consistency among readers. ACR TI-RADS has larger size thresholds for biopsy recommendations, along with options for active surveillance for smaller nodules with suspicious features, striking a balance between identifying the clinically significant cancers and the risk of overtreating. Strong evidence supports ACR TI-RADS as the preferred system for risk stratification of thyroid nodules.

Luyao Shen, MD
Stanford University School of Medicine
Stanford, CA
lyshen@stanford.edu

Version of record: Mar 22, 2023

The author declares that there are no disclosures relevant to the subject matter of this article.

doi.org/10.2214/AJR.23.29094

Provenance and review: Solicited; not externally peer reviewed.

References

1. 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
2. Kwak JY, Han KH, Yoon JH, et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. *Radiology* 2011; 260:892–899
3. Yang R, Zou X, Zeng H, Zhao Y, Ma X. Comparison of diagnostic performance of five different ultrasound TI-RADS classification guidelines for thyroid nodules. *Front Oncol* 2020; 10:598225