Performance of Five Ultrasound Risk Stratification Systems in Selecting Thyroid Nodules for FNA

Marco Castellana,[1] Carlo Castellana,[2] Giorgio Treglia,[3,4] Francesco Giorgino,[1] Luca Giovanella,[3,5] Gilles Russ,[6,7] and Pierpaolo Trimboli[3,8]

[1]Department of Emergency and Organ Transplantation, Section of Internal Medicine, Endocrinology, Andrology and Metabolic Diseases, University of Bari Aldo Moro, Bari, Italy; [2]University of Bari Aldo Moro, Bari, Italy; [3]Clinic for Nuclear Medicine and Competence Center for Thyroid Diseases, Imaging Institute of Southern Switzerland, Ente Ospedaliero Cantonale, Bellinzona, Switzerland; [4]Department of Nuclear Medicine and Molecular Imaging, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; [5]Medical School, University of Zurich, Zurich, Switzerland; [6]Thyroid Imaging and Cytopathology Centre, Paris, France; [7]Thyroid and Endocrine Tumors Unit, La Pitié Salpêtrière Hospital, Sorbonne University, Paris, France; and [8]Faculty of Biomedical Sciences, Università della Svizzera Italiana (USI), Lugano, Switzerland

ORCiD numbers: 0000-0002-1175-8998 (M. Castellana); 0000-0001-9808-780X (G. Treglia); 0000-0001-7372-2678 (F. Giorgino); 0000-0003-0230-0974 (L. Giovanella); 0000-0001-8036-6050 (G. Russ); 0000-0002-2125-4937 (P. Trimboli).

Context. Ultrasound (US) risk stratification systems (RSSs) have been developed to reduce the number of unnecessary fine-needle aspiration procedures (FNA) in patients with thyroid nodules.

Objective. We conducted a systematic review and meta-analysis evaluating the ability of the 5 most common US RSSs for the appropriate selection of thyroid nodules for FNA.

Data sources. This systematic review and meta-analysis was registered on PROSPERO (CRD42019131771). PubMed, CENTRAL, Scopus, and Web of Science were searched until March 2019.

Study selection. Original articles reporting data on the performance of AACE/ACE/AME, ACR TI-RADS, ATA, EU-TIRADS, and K-TIRADS were included.

Data extraction. The number of nodules classified as true negative, true positive, false negative, and false positive was extracted. Summary operating points were estimated using a random-effects model. Interobserver agreement was also assessed.

Data synthesis. Twelve studies evaluating 18 750 thyroid nodules were included. Participants were adult outpatients with thyroid nodules submitted to either FNA or core-needle biopsy or surgery and with available US images. The final diagnosis for malignant nodules was generally based on histology, while cytology was used for benign nodules. Diagnostic odds ratio (DOR) ranged from 2.2 to 4.9. A head-to-head comparison showed a higher relative DOR for ACR-TIRADS versus ATA ($P = .002$) or K-TIRADS ($P = .002$), due to a higher relative likelihood ratio for positive results.

Conclusions. The present meta-analysis found a higher performance of ACR TI-RADS in selecting thyroid nodules for FNA. However, the comparison across the most common US RSSs

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA

Received 4 July 2019. Accepted 5 November 2019.
First Published Online 6 November 2019.
Corrected and Typeset 30 March 2020.

was limited by the data available. Further studies are needed to confirm this finding.
(*J Clin Endocrinol Metab* 105: 1659–1669, 2020)

Key Words: thyroid nodule, ultrasound, systematic review, meta-analysis, diagnostic performance

Ultrasound (US) is the first-line imaging modality for malignancy risk assessment of thyroid nodules. Specific US features, such as hypoechogenicity, taller-than-wide shape on transverse view, irregular margins, microcalcifications, and extrathyroidal extension, are recognized to be associated with cancer (1). At the same time, using each individual feature as a standalone diagnostic parameter is associated to inter- and intra-operator variability (2). To mitigate these limitations, several US risk stratification systems (RSSs) have been developed to stratify the malignancy risk of a nodule and then suggest the need for fine-needle aspiration (FNA). Among these, 2 were included in clinical guidelines for the diagnosis and treatment of thyroid nodule and carcinoma (3,4), while the remaining ones were purely radiological recommendations (5–7). Three to six category scales were proposed, and an estimated risk of malignancy was assigned to each class.

Following the advent of these systems, several papers attempted to compare their performance. There, 2 specific outcomes were assessed: the risk of malignancy of each category and the performance in indicating FNA. The results of these studies have been heterogeneous, thus limiting the applicability of their findings to the clinical practice. Moreover, most of these studies had a retrospective design. Also, they enrolled nodules previously submitted to FNA during clinical practice and whose indication had not been based on these systems. Consequently, these studies were affected by a significant selection bias, which in turn impacted on the prevalence of malignancy (8–20). Indeed, it is well known that the performance of a diagnostic test depends on the frequency of the event (ie, disease) in the enrolled sample (21). As a proof, a significant difference in cancer rate was found in these studies, and as a consequence, a significant discrepancy was observed in terms of US RSSs performance. Given that US RSSs are diagnostic tests conceived for selecting thyroid nodules for FNA, we raise the question whether these systems are really comparable given the different methodologies of the published reports. Simply pooling the findings of prior studies would be associated with a significant bias. To overcome these limitations, summary operating measures assumed to be independent of the disease prevalence should be used. These include diagnostic odds ratio (DOR) and the likelihood ratio for positive results (LR+) and negative results (LR–) (21,22). The comparison of different US RSSs would then rely on relative measures, as relative DOR (RDOR), relative LR+ (RLR+), and relative LR– (RL–).

The present study aimed to reach information on this topic to reduce/delete the significant limitations of studies available in the literature. Then, here we planned (i) a systematic review of studies reporting the performance of the 5 most common US RSSs in selecting thyroid nodules for FNA; (ii) a meta-analysis of available data to evaluate the diagnostic performance of US RSSs; and (iii) a comparison of US RSSs. The primary outcomes were the DOR of each system taken independently and the RDOR when head-to-head comparison was feasible. The secondary outcomes were to evaluate sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), LR+, LR–, and interobserver agreement.

Materials and Methods

The systematic review was registered on PROSPERO (registration number CRD42019131771) and performed in accordance with the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) (23,24).

Search strategy

A 6-step search strategy was planned. First, sentinel studies were searched in PubMed. Second, keywords and MeSH terms were identified in PubMed. Third, to test the strategy, the terms "AACE/ACE/AME," "ACR TI-RADS," "EU-TIRADS," "K-TIRADS," and "ATA" were searched in PubMed (the full strategy can be found on PROSPERO). Fourth, PubMed, CENTRAL, Scopus, and Web of Science were searched. Fifth, studies reporting the diagnostic performance of at least 1 of the US RSSs in thyroid nodules were included if meeting both of the following criteria: (i) the diagnosis of benign nodules was based either on histology or core-needle biopsy (CNB) or cytology, and (ii) the diagnosis of malignant nodules was not based on cytology only. The latter criterion was adopted because although cytological diagnosis is reliable for papillary thyroid cancer, this is not true for follicular thyroid cancer (FTC), which is cytologically indistinguishable from its benign counterpart (follicular adenoma), or medullary thyroid cancer, which is missed by cytology in up to 50% of cases (25,26). Studies focusing on pediatric patients or specific subgroups of thyroid nodules (ie, indeterminate), as well as studies using cytology as the only reference standard for both malignant and benign nodules were excluded. Finally, references of included

studies were screened for additional papers. The last search was performed on March 25, 2019. No language or time restriction was adopted. Two investigators (MC, PT) independently and in duplicate searched papers, screened titles and abstracts of the retrieved articles, reviewed the full-texts, and selected articles for their inclusion.

Data extraction

The following information was extracted independently and in duplicate by 2 investigators (MC, PT) in a piloted form: (i) general information on the study (author, year of publication, country, study type, number of patients, number of nodules, final diagnosis, population); (ii) reference standard; (iii) number of nodules classified as true negative, true positive, false negative, and false positive; and (iv) interobserver agreement. Indication to FNA was the index test. A benign nodule was classified as true negative if FNA was not indicated by the specific US RSS. A benign nodule was classified as false positive if FNA was indicated by the specific US RSS. A malignant nodule was classified as true positive if FNA was indicated by the specific US RSS. A malignant nodule was classified as false negative if FNA was not indicated by the specific US RSS. The main paper and supplementary data were searched; if data were missing, authors were contacted via email. Data were cross-checked, and any discrepancy was discussed.

Study quality assessment

The risk of bias of included studies was assessed independently by 2 reviewers (MC, PT) through the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool for the following aspects: patient selection, index test, reference standard, and flow and timing. Risk of bias and concerns about applicability were rated as low, high, or unclear (27). Data presentation was arranged using RevMan 5.3 (Cochrane Collaboration, 2014).

Data analysis

The characteristics of included studies were summarized. Then, separate analyses were performed according to the following steps. First, a diagnostic performance meta-analysis in selecting nodules for FNA or not was carried out. For each US RSS, we plotted estimates of sensitivity and specificity on coupled forest plots. Summary operating points including sensitivity, specificity, NPV, PPV, LR+, LR–, and DOR, with 95% confidence intervals (95%CI), were estimated. DOR provides a single measure of test performance; it is equal to LR+/LR– and corresponds to the odds of the FNA being indicated in a malignant nodule compared to the odds of the FNA being indicated in a benign one. The value ranges from zero to infinity, with higher values indicating higher performance. LR+ is the likelihood for an US RSS that the FNA is indicated for a malignant nodule compared to the likelihood for a benign one. A LR+ greater than 10 means strong evidence; between 5 and 10, moderate evidence; and less than 5, weak evidence. LR– is the likelihood for an US RSS that the FNA is not indicated for a malignant nodule compared to the likelihood for a benign one. A LR- less than 0.1 means strong evidence; between 0.1 and 0.2, moderate evidence; and higher than 0.2, weak evidence. A bivariate random-effects model was used for the pooled analysis of sensitivity and specificity; a random-effects

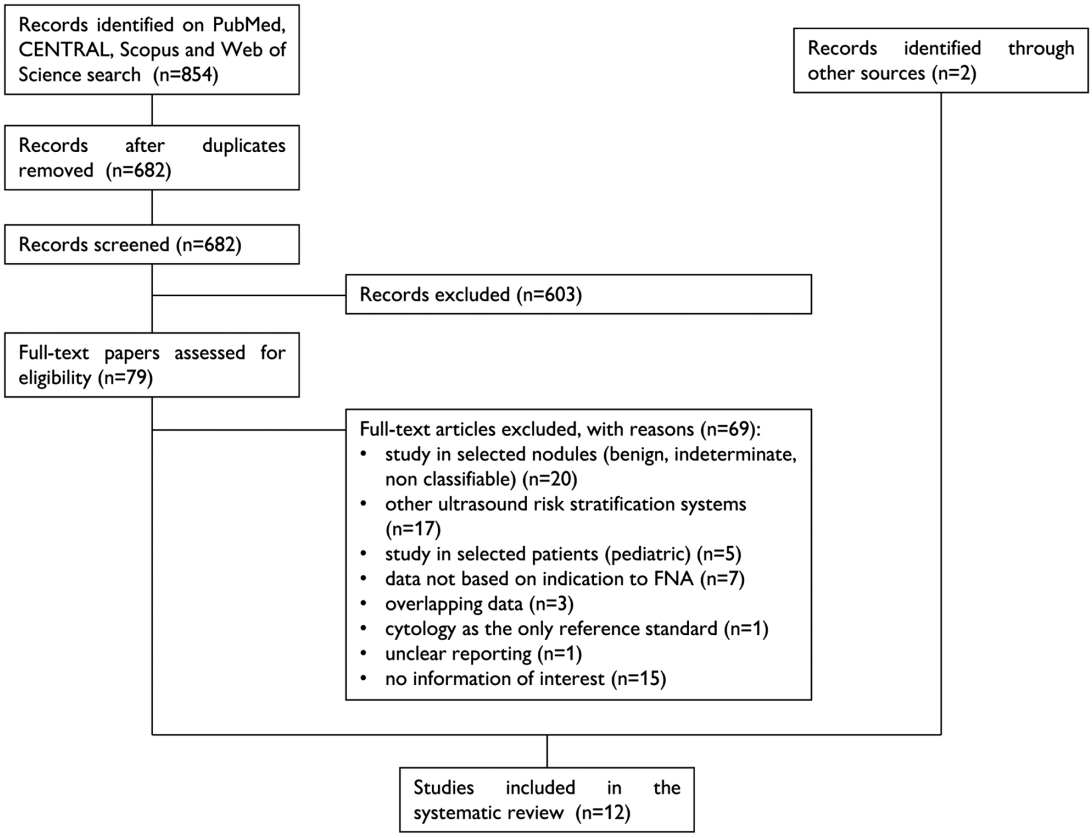

Figure 1. Flow chart of the systematic review.

Table 1. Characteristics of included studies

First Author, year	Country	Study design	Thyroid nodules (*n*)	Selection criteria of included study	AACE/ ACE/AME	ACR-TIRADS	ATA	EU-TIRADS	K-TIRADS	Interobserver agreement
Negro, 2017 (8)	Italy	RCS	629	FNA. Nondiagnostic and indeterminate FNA were excluded.	x					
Yoon, 2017 (9)	Korea	RCS	4696	TN measuring 10–19 mm. FNA. Nondiagnostic, indeterminate, or suspicious FNA were excluded.			x			
Ha, 2018 (10)	Korea	RCS	902	TN > 5 mm. FNA. Nondiagnostic and indeterminate FNA were excluded.		x	x		x	
Ha, 2018 (11)	Korea	RCS	2000	TN ≥ 10 mm. FNA or CNB	x	x	x		x	
Middleton, 2018 (12)	USA	RCS	3422	FNA. Nondiagnostic and indeterminate FNA were excluded.		x	x		x	
Persichetti, 2018 (13)	Italy	PCS[a]	987	FNA. Nondiagnostic, benign, and indeterminate with single FNA and suspicious without following surgery were excluded.	x		x			
Xu, 2018 (14)	China	RCS	2465	FNA. Nondiagnostic, indeterminate, and suspicious were excluded.		x		x	x	
Grani, 2019 (15)	Italy	PCS	502	FNA. Nondiagnostic and indeterminate FNA were excluded.	x	x	x	x	x	
Mohammadi, 2019 (16)	Canada	RCS	424	FNA			x			x
Ruan, 2019 (17)	China	RCS	1001	FNA. Nondiagnostic, indeterminate. and suspicious FNA were excluded.		x	x			
Trimboli, 2019 (18)	France, Switzerland and UK	RCS	1058	TN > 5mm. Histological diagnosis.				x		
Wu, 2019 (19)	China	NR	664	FNA. Nondiagnostic, indeterminate, and suspicious FNA were excluded.		x	x			

Abbreviations: CNB, core needle biopsy; FNA, fine-needle aspiration cytology; NR, not reported; PCS, prospective cohort study; RCS retrospective cohort study; TN, thyroid nodule; x, retrieved data.
[a]The authors described their study as prospective even though the design seems to be retrospective.

model was used for the pooled analysis of the remaining metrics (28). Second, a head-to-head comparison on the accuracy of US RSSs was performed if at least 5 studies were available. Two systems were included in each comparison, and they were classified arbitrarily as "US RSS A" and "US RRS B." The significance of the differences between US RSSs was assessed with RDOR, RLR+, and RLR–, with 95%CI. The value ranges from zero to infinity, and if its 95%CI does not include the value 1, there is a statistically significant difference between the 2 systems (28,29). All analyses were performed on a per-lesion basis and carried out using RevMan 5.3 (the Cochrane Collaboration) and R 3.5.2 (Core Team). Heterogeneity between studies was assessed by using I^2, with 50% or higher values regarded as high heterogeneity. Publication bias was not evaluated, because of uncertainty about the determinants for diagnostic accuracy studies and the inadequacy of tests for detecting funnel plot asymmetry (29). A $P < .05$ was regarded as significant.

Results

A total of 854 papers were found, of which 128 were on PubMed; 586, on Scopus; 90, on Web of Science; and 50, on CENTRAL. After removal of 172 duplicates, 682 articles were analyzed for title and abstract; 603 records were excluded (reviews, guidelines, not within the field of the review, studies not in humans). The remaining 79 papers were retrieved in full-text, and 10 articles were finally included in the systematic review (Fig. 1) (8,10–17,19). Additionally, one article was added after screening the references of these papers, and another one was added from a personal database (9,18).

Qualitative analysis (systematic review)

The characteristics of the included articles are summarized in Table 1. The papers were published between 2017 and 2019 and had sample sizes ranging from 424 to 4696 thyroid nodules. Participants were adult outpatients who had undergone either thyroid nodule FNA, CNB, or surgery and had US images available. Thyroid nodules diagnosed as indeterminate on FNA were generally excluded, unless a final diagnosis was met on pathology. One study included only patients for which histological diagnosis and US images were available (18). Nine studies were retrospective, and 2 used prospective cohorts; the design was not clearly stated in 1 paper (19). Three studies were carried in China, 3 in South Korea, 3 in Italy, 1 in the United States, 1 in Canada, and 1 in France, Switzerland, and the United Kingdom. Four studies assessed AACE/ACE/AME's US RSS; 7, ACR TI-RADS; 9, ATA's US RSS; 3, EU-TIRADS; and 5, K-TIRADS. The reference standard for malignant and benign diagnosis is reported in Table 2. The prevalence of malignancy ranged from 4% to 54% (8,19). Overall, 4378 malignant and 14 372 benign nodules were included in the present review.

Quantitative analysis (meta-analysis)

The forest plot of the sensitivity and specificity of each US RSS in selecting thyroid nodules for FNA is shown in Fig. 2. The pooled sensitivities ranged from 54% to 87%; specificities, from 28% to 64%; PPVs, from 17% to 43%; and NPVs, from 81% to 93%. Since these summary operating points are influenced by the prevalence of the disease in the population tested, we estimated the following parameters, which are independent of disease prevalence and thus characteristics of the specific US RSSs. The pooled LR+ ranged from 1.2 to 1.9; LR–, from 0.4 to 0.6; and DOR, from 2.2 to 4.9. A high heterogeneity was found for all the outcomes (Table 3).

Table 2. Reference standard

First Author, year	Reference standard for malignant lesions	Reference standard for benign lesions
Negro, 2017 (8)	Histology (100%)	Cytology (100%)
Yoon, 2017 (9)	Histology (88%), cytology (12%)	Histology (4%), single (91%) or repeated cytology (5%)
Ha, 2018 (10)	Histology (72%), cytology or CNB (28%)	Histology (6%), cytology (94%)
Ha, 2018 (11)	Histology (99%), cytology or CNB (1%)	Histology (15%), repeated cytology or CNB (25%), cytology or CNB and follow-up (60%)
Middleton, 2018 (12)	Histology (86%), cytology (14%)	Histology (NR), cytology (NR)
Persichetti, 2018 (13)	Histology (100%)	Histology (13%), repeated cytology (87%)
Xu, 2018 (14)	Histology (100%)	Histology (35%), cytology and follow-up (65%)
Grani, 2019 (15)	Histology (94%), cytology (6%)	Histology (NR), cytology and follow-up (NR)
Mohammadi, 2019 (16)	Histology (NR), cytology (NR)	Histology (NR), cytology (NR)
Ruan, 2019 (17)	Histology (92%), cytology (8%)	Histology (36%), cytology (64%)
Trimboli, 2019 (18)	Histology (100%)	Histology (100%)
Wu, 2019 (19)	Histology (69%), cytology or CNB (31%)	Histology (2%), repeated cytology or CNB (27%), cytology or CNB and follow-up (71%)

Abbreviations: CNB, core needle biopsy; NR, not reported.

AACE/ACE/AME

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Grani, 2019 [15]	31	296	5	170	0.86 [0.71, 0.95]	0.36 [0.32, 0.41]
Ha, 2018 [11]	365	649	89	897	0.80 [0.76, 0.84]	0.58 [0.56, 0.60]
Negro, 2017 [8]	19	228	6	376	0.76 [0.55, 0.91]	0.62 [0.58, 0.66]
Persichetti, 2018 [13]	85	451	71	380	0.54 [0.46, 0.62]	0.46 [0.42, 0.49]

ACR-TIRADS

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Grani, 2019 [15]	30	204	6	262	0.83 [0.67, 0.94]	0.56 [0.52, 0.61]
Ha, 2018 [10]	217	200	49	436	0.82 [0.76, 0.86]	0.69 [0.65, 0.72]
Ha, 2018 [11]	339	505	115	1041	0.75 [0.70, 0.79]	0.67 [0.65, 0.70]
Middleton, 2018 [12]	240	1447	112	1623	0.68 [0.63, 0.73]	0.53 [0.51, 0.55]
Ruan, 2019 [17]	209	142	183	467	0.53 [0.48, 0.58]	0.77 [0.73, 0.80]
Wu, 2019 [19]	327	155	32	150	0.91 [0.88, 0.94]	0.49 [0.43, 0.55]
Xu, 2018 [14]	494	427	511	1033	0.49 [0.46, 0.52]	0.71 [0.68, 0.73]

ATA

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Grani, 2019 [15]	27	255	9	211	0.75 [0.58, 0.88]	0.45 [0.41, 0.50]
Ha, 2018 [10]	239	350	27	286	0.90 [0.86, 0.93]	0.45 [0.41, 0.49]
Ha, 2018 [11]	407	1033	47	513	0.90 [0.86, 0.92]	0.33 [0.31, 0.36]
Middleton, 2018 [12]	235	2054	72	584	0.77 [0.71, 0.81]	0.22 [0.21, 0.24]
Mohammadi, 2019 [16]	31	347	1	45	0.97 [0.84, 1.00]	0.11 [0.08, 0.15]
Persichetti, 2018 [13]	93	564	63	267	0.60 [0.51, 0.67]	0.32 [0.29, 0.35]
Ruan, 2019 [17]	223	356	169	253	0.57 [0.52, 0.62]	0.42 [0.38, 0.46]
Wu, 2019 [19]	348	234	11	71	0.97 [0.95, 0.98]	0.23 [0.19, 0.28]
Yoon, 2017 [9]	999	2165	45	1487	0.96 [0.94, 0.97]	0.41 [0.39, 0.42]

EU-TIRADS

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Grani, 2019 [15]	31	317	5	149	0.86 [0.71, 0.95]	0.32 [0.28, 0.36]
Trimboli, 2019 [18]	171	297	86	504	0.67 [0.60, 0.72]	0.63 [0.59, 0.66]
Xu, 2018 [14]	498	662	507	798	0.50 [0.46, 0.53]	0.55 [0.52, 0.57]

K-TIRADS

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Grani, 2019 [15]	33	383	3	83	0.92 [0.78, 0.98]	0.18 [0.14, 0.22]
Ha, 2018 [10]	244	396	22	240	0.92 [0.88, 0.95]	0.38 [0.34, 0.42]
Ha, 2018 [11]	429	1138	25	408	0.94 [0.92, 0.96]	0.26 [0.24, 0.29]
Middleton, 2018 [12]	267	2362	72	587	0.79 [0.74, 0.83]	0.20 [0.18, 0.21]
Xu, 2018 [14]	592	791	413	669	0.59 [0.56, 0.62]	0.46 [0.43, 0.48]

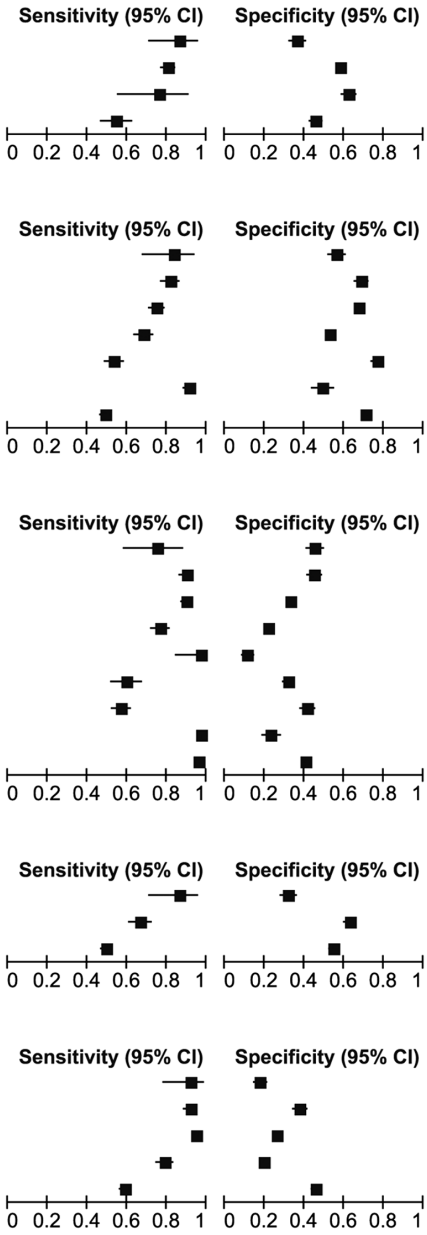

Figure 2. Forest plot of sensitivity and specificity of each ultrasound risk stratification system in selecting thyroid nodules for FNA. Abbreviations: FN, false negative; FP, false positive; TN, true negative; TP, true positive.

Second, we made compared the accuracy of those US RSSs for which data from at least 5 studies were available (ATA, ACR TI-RADS, and K-TIRADS). When we considered only those studies assessing simultaneously all evaluated US RSSs through a head-to-head comparison, a higher RDOR was found for ACR TI-RADS versus ATA (P =.002) or K-TIRADS (P = .002) (Table 4; Fig. 3). A higher RLR+ was found for ACR TI-RADS versus ATA (P < .001) or K-TIRADS (P < .001) as well as for ATA versus K-TIRADS (P = .048) (Table 5). No difference in RLR– was found (Table 6).

Finally, only 1 study assessed interobserver agreement in recommending FNA. Mohammadi et al reported a moderate agreement for ATA's US RSS between the 2 radiologists participating in the study (16).

Study quality assessment

The risk of bias of the included studies is shown in a digital research materials repository (24). Overall, we found a low risk of bias: in most studies patients included were consecutive ones who underwent US and had a final diagnosis in a specific period; the classification according to US RSSs was conducted before the final diagnosis; in retrospective studies, researchers were blinded to the final diagnosis; or, in Grani and in Yoon, features recorded during US examination before FNA

Table 3. Summary estimates of the accuracy of each ultrasound risk stratification system in selecting thyroid nodules for FNA

Ultrasound risk stratification system	Number of nodules (number of studies)	Prevalence of malignancy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	Positive predictive value (95%CI)	Negative predictive value (95%CI)	Likelihood ratio for positive results (LR+) (95% CI)	Likelihood ratio for negative results (LR-) (95% CI)	Diagnostic odds ratio (DOR)
AACE/ACE/AME	4118 (4)	12 (3–21)	74 (71–78)	53 (51–55)	17 (4–30)	93 (87–98)	1.5 (1.1–2.1)	0.5 (0.2–1.0)	3.1 (1.0–9.4)
ACR TI-RADS	12 996 (7)	29 (17–41)	74 (61–83)	64 (56–70)	43 (25–61)	84 (77–93)	1.9 (1.6–2.3)	0.4 (0.3–0.6)	4.9 (3.1–7.7)
ATA	14 121 (9)	23 (16–30)	87 (75–94)	31 (24–40)	27(17–36)	88 (83–93)	1.2 (1.0–1.4)	0.4 (0.2–0.7)	3.1 (1.3–7.1)
EU-TIRADS	4025 (3)	24 (4–44)	54 (51–57)	53 (51–55)	29 (7–52)	81 (60–100)	1.4 (1.0–1.8)	0.6 (0.4–1.0)	2.2 (0.9–5.1)
K-TIRADS	9157 (5)	22 (10–34)	86 (73–94)	28 (20–38)	25 (12–39)	87 (75–99)	1.2 (1.0–1.4)	0.5 (0.2–1.0)	2.5 (1.1–5.5)

were used (9,15). We rated the reference standard bias as high, due to cytology being generally adopted for benign nodules. We rated the flow and timing bias as low, given that thyroid cancer is a chronic condition. The only exception to the previous statements was 2 studies in which the patient selection risk of bias was rated as unclear due to missing information on consecutive or random enrollment (8,9). In 1 study, histology was adopted as the reference standard, and thus the corresponding item was rated as low (18). Finally, 4 studies excluded nodules depending on their size or composition, and thus the patient selection applicability concerns item was rated as high (9–11,18).

Discussion

The aim of this systematic review was to identify the best available evidence on the diagnostic performance of the five most common US classification systems in the indication for FNA. To our knowledge, this is the first systematic review and meta-analysis on this topic, allowing studies evaluating populations with a different prevalence of malignancy to be interpreted together. An extensive database search was performed without time or language restrictions, and inclusion criteria were defined prior to the database search. Twelve studies were found, evaluating 4378 malignant and 14 372 benign thyroid nodules. There were sparse data on AACE/ACE/AME and EU-TIRADS. On the contrary, a higher number of studies evaluating ACR TI-RADS, ATA, and K-TIRADS was found.

Sensitivities of US RSSs ranged from 54% to 87%; specificities, 28% to 64%; PPV, 17% to 43%; and NPV, 81% to 93%. LR+, LR–, and DOR ranged between 1.2 and 1.9, 0.4 and 0.6, and 2.2 and 4.9, respectively (Table 3). Thus, a wide interval was observed, with the notable exception of NPV, which was high in almost every report. The results of LR+ and LR– showed weak evidence across all US RSSs for the effectiveness to correctly select thyroid nodules for FNA. Also, a high heterogeneity was estimated for all summary operating points. This was expected for those parameters known to be influenced by the prevalence of the disease in the population tested (ie, PPV, NPV), but not for other parameters (ie, LR+, LR–, DOR) known to be characteristics of the specific US RSSs. In this case, it is reasonably attributable to the US being an operator-dependent imaging modality. To overcome the previously noted limitations, head-to-head comparisons based on relative measurements (ie, RLR+, RLR–, RDOR) were performed. A higher performance for ACR TI-RADS compared to ATA or K-TIRADS was found. The

Table 4. Head-to-head comparison of DOR of ultrasound risk stratification systems for selecting thyroid nodules for FNA

US RSS A	US RSS B	Number of nodules (number of studies)	DOR of US RSS A	DOR of US RSS B	RDOR	*P*
ACR TI-RADS	ATA	8491 (6)	5.6 (3.4–9.0)	2.9 (1.3–6.5)	1.9 (1.3–2.9)	**0.002**
ATA	K-TIRADS	6692 (4)	2.9 (1.0–8.2)	3.1 (0.9–10.7)	0.9 (0.8–1.1)	0.552
ACR-TIRADS	K-TIRADS	9291 (5)	4.5 (2.5–7.9)	2.5 (1.1–5.6)	1.8 (1.2–2.6)	**0.002**

US RSS A and US RSS B represent the two systems considered for the specific comparison. RDOR value ranges from zero to infinity; if the 95%CI does not include the value 1, there is a statistically significant difference between the two systems.
Abbreviations: 95%CI, 95% confidence interval; DOR, diagnostic odds ratio; RDOR, relative diagnostic odds ratio; US RSS, ultrasound risk stratification system.

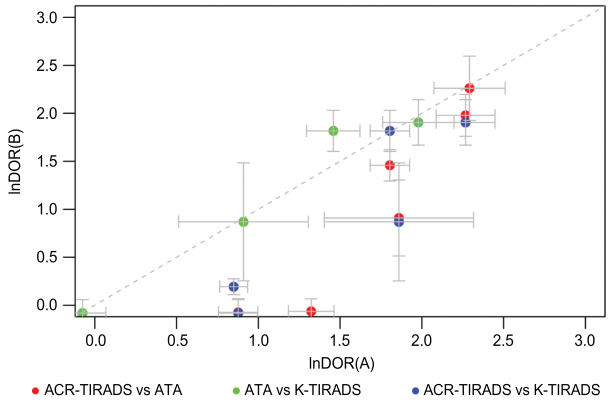

Figure 3. Head-to-head comparison of DOR of ultrasound risk stratification systems for selecting thyroid nodules for FNA. On the *x* and *y* axis, the natural logarithm of DOR for a specific US RSSs for each study is reported according to Table 4. Relative diagnostic odds ratios (RDOR) with 95% confidence intervals for each study according to a specific head-to-head comparison are plotted. If the 95%CI of the RDOR does not include the dashed line (corresponding to a value of 1), there is a statistically significant difference between the 2 systems for the specific study. In example, in the top right corner the RDOR of a study comparing ACR TI-RADS versus ATA is plotted; since its 95% CIincludes the dashed line, there is no difference between the 2 US RSSs. Abbreviations: DOR, diagnostic odds ratio.

discriminative power was related to a higher ability of ACR TI-RADS to select malignant nodules for FNA (ie, higher RLR+), while no difference was found for benign nodules (ie, similar RLR–). Finally, there was limited evidence for differences in the interobserver agreement among US RSSs, with only Mohammadi et al reporting a moderate agreement for ATA's US RSS between the 2 radiologists participating in the study (16). However, a recent paper found substantial to near-perfect agreement for all the US RSSs included in our study (30).

Four main explanations can be found for the results. First, despite similarities, definitions differ among US RSSs. For example, the EU-TIRADS gives a restrictive description of microcalcifications, distinguishing them from other hyperechoic foci, while the other systems have less stringent definitions. Second, nodule risk's classification is performed according to specific criteria in each US RSS. For example, microcalcifications confer "per se'" a high-risk score according to AACE/ACE/AME and EU-TIRADS, while for the remaining systems, additional features are required. Third, the estimated risk of malignancy assigned to each class differed between US RSSs. Regarding the intermediate risk categories, EU-TIRADS, ATA, AACE/ACE/AME, and ACR TI-RADS are quite close to one another (6%–17%, 10%–20%, 5%–15%, and 5%–20%, respectively). On the contrary, the range of K-TIRADS is wider (15%–50%). For high-risk categories, the estimated risks are close for K-TIRADS, ATA, and AACE/ACE/AME (>60%, >70%–90%, and 50%–90%, respectively), but very different for EU-TIRADS and ACR TI-RADS (26%–87% and >20%, respectively). Finally, the cut-offs for FNA also vary from one system to another. While they all agree on the 10 mm cut-off for highly suspicious nodules, the cut-off for intermediate ones is 20 mm in AACE/ACE/AME, 15 mm in EU-TIRADS and ACR TI-RADS, and 10 mm in K-TIRADS and ATA system. As these represent a substantial portion of all nodules, differences in cut-offs will significantly modify the diagnostic values (3–7).

What could be the reason of the ACR TI-RADS's better performance? Since studies included in the head-to head comparisons assessed the performance of US RSSs within the same population, the results of our meta-analysis showing a higher performance for ACR TI-RADS should not be impaired by recruitment, cancer prevalence, or operator biases. This system does not particularly excel in its sensitivity, NPV, PPV, or LR– ratio. However, its specificity is significantly above that of the other systems, explaining the higher LR+ and DOR. This can't be explained by the size cut-offs for FNA in intermediate- and high-risk-nodules, given that it is similar to that of the other US RSSs. However, if we refer to the US RSS itself, it is very likely that fewer nodules are categorized in the intermediate- or high-suspect categories, compared to the other systems. The main reason for this could be that nodules with regular

Table 5. Head-to-head comparison of LR+ of ultrasound risk stratification systems for selecting thyroid nodules for FNA

US RSS A	US RSS B	Number of nodules (number of studies)	LR+ of US RSS A	LR+ of US RSS B	RLR+	*P*
ACR TI-RADS	ATA	8491 (6)	2.0 (1.7–2.4)	1.2 (1.1–1.5)	1.6 (1.4–1.9)	**<0.001**
ATA	K-TIRADS	6692 (4)	1.3 (1.1–1.6)	1.2 (1.0–1.4)	1.1 (1.0–1.2)	**0.048**
ACR-TIRADS	K-TIRADS	9291 (5)	1.9 (1.6–2.4)	1.2 (1.0–1.4)	1.7 (1.5–1.8)	**<0.001**

US RSS A and US RSS B represent the two systems considered for the specific comparison. RLR+ value ranges from zero to infinity; if the 95% confidence interval of does not include the value 1, there is a statistically significant difference between the two systems.
Abbreviations: LR+, likelihood ratio for positive results; RLR+, relative likelihood ratio for positive results; US RSS, ultrasound risk stratification system.

Table 6. Head-to-head comparison of LR- of ultrasound risk stratification systems for selecting thyroid nodules for FNA

US RSS A	US RSS B	Number of nodules (number of studies)	LR– of US RSS A	LR- of US RSS B	RLR-	*P*
ACR TI-RADS	ATA	8491 (6)	0.4 (0.3–0.5)	0.4 (0.2–0.8)	0.8 (0.6–1.2)	0.338
ATA	K-TIRADS	6692 (4)	0.5 (0.2–0.9)	0.5 (0.2–0.9)	1.2 (1.0–1.4)	0.114
ACR-TIRADS	K-TIRADS	9291 (5)	0.4 (0.3–0.7)	0.4 (0.2–0.9)	0.9 (0.6–1.4)	0.673

US RSS A and US RSS B represent the 2 systems considered for the specific comparison. RLR– value ranges from zero to infinity; if the 95% confidence interval of does not include the value 1, there is a statistically significant difference between the two systems.
Abbreviations: LR–, likelihood ratio for negative results; RLR–, relative likelihood ratio for negative results; US RSS, ultrasound risk stratification system.

shape and margins, mild hypoechogenicity, and mixed composition will be classified as ACR TI-RADS 3, while being categorized as EU-TIRADS 4 and AACE/ACE/AME class 2 (intermediate risk). They cannot be classified in the ATA system, which is reportedly characterized by a relevant number of not-classifiable lesions (31–33). In the K-TIRADS system, they would be categorized as K-TIRADS 3, as in the ACR TI-RADS. However, the K-TIRADS itself is probably hampered by the low FNA cut-off chosen for nodules at intermediate risk (10 mm). As these intermediate risk nodules are very frequent, this could explain the advantage of the ACR TI-RADS over the other systems (3–7).

Some possible drawbacks of adopting US RSSs should be considered. First, as stated, their broad applicability can be limited by a variable number of nodules in which these systems cannot be used, since some nodules cannot be classified. Second, the assessment of each nodule can be particularly time consuming, especially in those patients with multinodular goiters. Indeed, the pattern-based systems (eg, ATA and K-TIRADS) are less time-consuming than score-based systems (eg, ACR TI-RADS). Third, all US RSSs emphasize the hypoechoic nature of malignancies, but this is primarily specific for the papillary thyroid cancer. Possibly, the systems may miss the small FTCs, since they often present as isoechoic, not-calcified, round lesions (25). It is worth noting that the risk of distant metastasis in patients with FTC increases for tumor larger than 20 mm (34). Also, the FTCs size reported in the literature was generally higher than any dimensional cut-off proposed in the US RSSs, so it is possible that no different results could be met if FNA indication is based either on US RSS or solely on clinical judgment (35,36).

This review has several limitations. The first limitation relates to the design of included studies: a retrospective review and reclassification of nodules that have been submitted to FNA was performed in most of them, with possible selection bias. In some studies, US images were retrospectively reviewed, and this was a second limitation (10–12,16,18). Although the inter-exam agreement between real-time and retrospective US image interpretation for thyroid nodules was found to be equal or more than substantial, it is possible that an examination not carefully performed during clinical practice would lead to an unreliable reassessment (37). Third, diagnosis of both benign and malignant lesions was often based on cytology or CNB, with a possible reference standard bias. Lastly, there were not enough studies to perform an analysis on all the included US classification systems.

The advantages of adopting US RSSs in improving the selection of thyroid nodules is recognized, and several options are available in the literature. However, poor data have been available on the performance of these US RSSs in selecting thyroid nodules for FNA. To date, sparse data allowed us to compare some of them, and the main finding of this study was that ACR TI-RADS had the highest performance compared to ATA or K-TIRADS. Further prospective studies assessing all of

the most common US RSSs and adopting histology as standard of reference are needed.

Acknowledgments

Author Contributions: MC and PT conceived the meta-analysis. All authors contributed to the development of the selection criteria, the risk of bias assessment strategy, and data extraction criteria. MC and PT developed the search strategy, performed database search, acquired the data, analyzed the data, and drafted the manuscript. CC provided statistical expertise. All authors read, provided feedback, and approved the final manuscript.

Additional Information

Correspondence and Reprint Requests: Marco Castellana, MD, Department of Emergency and Organ Transplantation, Section of Internal Medicine, Endocrinology, Andrology and Metabolic Diseases, University of Bari Aldo Moro, Bari, Italy. E-mail: marco.castellana@uniba.it.

Disclosure Summary: The authors have nothing to disclose.

Data Availability: The data sets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

References

1. Hegedüs L. Clinical practice. The thyroid nodule. *N Engl J Med*. 2004;**351**(17):1764–1771.
2. Hoang JK, Middleton WD, Farjat AE, et al. Interobserver variability of sonographic features used in the American college of radiology thyroid imaging reporting and data system. *AJR Am J Roentgenol*. 2018;**211**(1):162–167.
3. Gharib H, Papini E, Garber JR, Duick DS, 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**(5):622–639.
4. 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):1–133.
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**(3):370–395.
6. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association Guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *Eur Thyroid J*. 2017;**6**(5):225–237.
7. 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**(5):587–595.
8. Negro R, Greco G, Colosimo E. Ultrasound risk categories for thyroid nodules and cytology results: a single institution's experience after the adoption of the 2016 update of medical guidelines by the American Association of Clinical Endocrinologists and Associazione Medici Endocrinologi. *J Thyroid Res*. 2017;**2017**:8135415.
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**(2):560–569.
10. Ha EJ, Na DG, Moon WJ, Lee YH, Choi N. Diagnostic performance of ultrasound-based risk-stratification systems for thyroid nodules: comparison of the 2015 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**(11):1532–1537.
11. 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**(3):893–900.
12. Middleton WD, Teefey SA, Reading CC, 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**(5):1148–1154.
13. 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**(4):1359–1368.
14. 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**(2):299–307.
15. Grani G, Lamartina L, Ascoli V, et al. Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the "Right" TIRADS. *J Clin Endocrinol Metab*. 2019;**104**(1):95–102.
16. Mohammadi M, Betel C, Burton KR, Higgins KM, Ghorab Z, Halperin IJ. Retrospective application of the 2015 American Thyroid Association Guidelines for ultrasound classification, biopsy indications, and follow-up imaging of thyroid nodules: can improved reporting decrease testing? *Can Assoc Radiol J*. 2019;**70**(1):68–73.
17. 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**(9):4871–4878.
18. Trimboli P, Ngu R, Royer B, et al. A multicentre validation study for the EU-TIRADS using histological diagnosis as a gold standard. *Clin Endocrinol (Oxf)*. 2019;**91**(2):340–347.
19. 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**(1):121–131.
20. Haneuse S. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Med Care*. 2016;**54**(4):e23–e29.
21. Eusebi P. Diagnostic accuracy measures. *Cerebrovasc Dis*. 2013;**36**(4):267–272.
22. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;**56**(11):1129–1135.
23. McInnes MDF, Moher D, Thombs BD, et al; the PRISMA-DTA Group. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA*. 2018;**319**(4):388–396.
24. Castellana M, Castellana C, Treglia G, et al. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. A systematic review and meta-analysis. Supplemental material R1. Fig Share Digital Repository 2019. https://doi.org/10.6084/m9.figshare.9878333.v1. https://figshare.com/articles/Castellana_Supplemental_material_R1/9878333). Deposited September 19, 2019.
25. Grani G, Lamartina L, Durante C, Filetti S, Cooper DS. Follicular thyroid cancer and Hürthle cell carcinoma: challenges in diagnosis, treatment, and clinical management. *Lancet Diabetes Endocrinol*. 2018;**6**(6):500–514.

26. Trimboli P, Treglia G, Guidobaldi L, et al. Detection rate of FNA cytology in medullary thyroid carcinoma: a meta-analysis. *Clin Endocrinol (Oxf)*. 2015;**82**(2):280–285.
27. Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;**155**(8):529–536.
28. European Network for Health Technology Assessment. Meta-analysis of diagnostic test accuracy studies. https://www.eunethta.eu/wp-content/uploads/2018/01/Meta-analysis-of-Diagnostic-Test-Accuracy-Studies_Guideline_Final-Nov-2014.pdf. Accessed March 01, 2019.
29. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11: Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9*. The Cochrane Collaboration. https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/DTA%20Handbook%20Chapter%2011%20201312.pdf. Published 2013.
30. Grani G, Lamartina L, Cantisani V, Maranghi M, Lucia P, Durante C. Interobserver agreement of various thyroid imaging reporting and data systems. *Endocr Connect*. 2018;**7**(1):1–7.
31. Rosario PW, da Silva AL, Nunes MS, Ribeiro Borges MA5 Mourão GF, Calsolari MR. Risk of malignancy in 1502 solid thyroid nodules >1 cm using the new ultrasonographic classification of the American Thyroid Association. *Endocrine*. 2017;**56**(2):442–445.
32. Lauria Pantano A, Maddaloni E, Briganti SI, et al. Differences between ATA, AACE/ACE/AME and ACR TI-RADS ultrasound classifications performance in identifying cytological high-risk thyroid nodules. *Eur J Endocrinol*. 2018;**178**(6):595–603.
33. Ahmadi S, Oyekunle T, Jiang X, et al. A direct comparison of the ATA and TI-RADS ultrasound scoring systems. *Endocr Pract*. 2019;**25**(5):413–422.
34. Machens A, Holzhausen HJ, Dralle H. The prognostic value of primary tumor size in papillary and follicular thyroid carcinoma. *Cancer*. 2005;**103**(11):2269–2273.
35. Lai X, Jiang Y, Zhang B, et al. Preoperative sonographic features of follicular thyroid carcinoma predict biological behavior: A retrospective study. *Medicine (Baltimore)*. 2018;**97**(41):e12814.
36. Kim H, Shin JH, Hahn SY, et al. Prediction of follicular thyroid carcinoma associated with distant metastasis in the preoperative and postoperative model. *Head Neck*. 2019;**41**(8):2507–2513.
37. Bae JM, Hahn SY, Shin JH, Ko EY. Inter-exam agreement and diagnostic performance of the Korean thyroid imaging reporting and data system for thyroid nodule assessment: Real-time versus static ultrasonography. *Eur J Radiol*. 2018;**98**:14–19.