Original Article

Comparison of diagnostic performance of the current score-based ultrasound risk stratification systems according to thyroid nodule size

Cai-Feng Si, Jing Yu, Yi-Yang Cui, Yuan-Jing Huang, Ke-Fei Cui, Chao Fu

Department of Ultrasound, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Contributions: (I) Conception and design: C Fu, KF Cui, CF Si; (II) Administrative support: KF Cui, J Yu; (III) Provision of study materials or patients: CF Si; (IV) Collection and assembly of data: CF Si; (V) Data analysis and interpretation: CF Si, YJ Huang, C Fu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Chao Fu, MM. Department of Ultrasound, The First Affiliated Hospital of Zhengzhou University, No. 1 Jianshe East Road, Erqi District, Zhengzhou 450052, China. Email: fuchao3380@163.com.

Background: The lack of standardization in risk stratification systems (RSSs) has led to uncertainty in selecting the most effective RSS for diagnosing malignancy risk in thyroid nodules. Therefore, the aim of this study was to compare the diagnostic performance of four current score-based RSSs according to thyroid nodule size, with the goal of determining the most effective RSS and aiding in clinical decision-making.

Methods: Between July 2013 and January 2019, a total of 2,667 consecutive patients presenting with 3,944 thyroid nodules were pathologically diagnosed after thyroidectomy and/or ultrasound (US)-guided fine-needle aspiration (FNA). These nodules were retrospectively dichotomized into two groups: small nodules (<1 cm) and large nodules (≥1 cm). The four RSSs were used to assign US categories, and the diagnostic performances were computed and compared based on the size of thyroid nodules, both before and after the application of size thresholds for biopsy.

Results: After thyroidectomy or biopsy, 1,781 (45.2%) thyroid nodules were found to be malignant. (I) After applying size thresholds for biopsy in ≥1 cm nodules, the highest specificity, accuracy, area under the curve (AUC) and the lowest FNA rate and unnecessary FNA rate were observed in the Artificial Intelligence-Thyroid Imaging Reporting And Data System (AI-TIRADS) (66.1%, 75.3%, 0.785, 55.1%, and 38.6%, respectively, P<0.05 for all). (II) Before applying size thresholds for biopsy in ≥1 cm nodules, the FNA rate and unnecessary FNA rate of the four RSSs were lower they were after the application of the size threshold: American College of Radiology Thyroid Imaging Reporting and Data System (ACR-TIRADS), 59.1% versus 61.4%, 39.8% versus 45.4%; AI-TIRADS, 52.3% versus 55.1%, 34.0% versus 38.6%; TIRADS issued by Kwak *et al.* (Kwak-TIRADS), 52.5% versus 76.1%, 34.4% versus 52.1%; Chinese Thyroid Imaging Reporting and Data System (C-TIRADS), 51.5% versus 66.2%, 34.4% versus 50.1% (P<0.05 for all). (III) The small nodules showed higher sensitivity and lower specificity than the large nodules (ACR-TIRADS, 97.7% versus 95.5%, 46.2% versus 62.5%; AI-TIRADS, 97.2% versus 92.7%, 49.9% versus 71.6%; Kwak-TIRADS, 97.2% versus 92.5%, 49.7% versus 71.3%; C-TIRADS, 94.2% versus 90.7%, 55.0% versus 71.8%, respectively, all P<0.05).

Conclusions: A potential effective strategy for managing large nodules in the current score-based RSSs could be to rely solely on US categories rather than size thresholds for biopsy. Additionally, the diagnostic performance of small nodules showed higher sensitivity and lower specificity compared to large nodules before applying size thresholds for biopsy. These findings suggest a possible new management strategy for large nodules and provide a basis for the managing small nodules.

Keywords: Ultrasonography; thyroid nodule; risk stratification system (RSS); fine-needle aspiration (FNA)

Submitted Feb 13, 2024. Accepted for publication Sep 18, 2024. Published online Nov 06, 2024.
doi: 10.21037/qims-24-282
View this article at: https://dx.doi.org/10.21037/qims-24-282

Introduction

Ultrasound (US) is the preferred imaging modality for the evaluation of thyroid nodules (1) and several international societies have proposed US-based risk stratification systems (RSSs) for thyroid nodules (2-6). Multiple studies have compared the diagnostic performance of various RSSs to determine which is most effective (7-9), providing a basis for selecting the optimal RSS in daily clinical practice.

Current RSSs have been dichotomized into a pattern-based RSS and a score-based RSS. The pattern-based RSS involves the recognition of a grouping of US features (3,6,10), whereas the score-based RSS proposed a different triage based on a quantitative scoring system which is summed up to a numeric score resulting in a final category. The score-based RSS is applicable to all of the nodules and has been shown to be practical and easy to apply (11-13). Meanwhile, the score-based RSSs has shown higher specificity, accuracy, positive predictive value, and area under the curve (AUC), and lower unnecessary fine-needle aspiration (FNA) rates (14). Despite the effectiveness of score-based RSSs, less information is available regarding the comparative diagnostic performance of current score-based RSSs, such as the Chinese Thyroid Imaging Reporting and Data System (C-TIRADS) (2), Kwak-TIRADS (which was issued by Kwak *et al.*) (15), the American College of Radiology Thyroid Imaging Reporting and Data System (ACR-TIRADS) (4), and Artificial Intelligence-Thyroid Imaging Reporting And Data System (AI-TIRADS) (which was a simplified version of ACR-TIRADS by artificial intelligence algorithm) (16).

Currently, when a thyroid nodule exhibits suspicious features, a 1 cm size threshold is commonly used to trigger a US-guided biopsy. As a result, the majority of patients enrolled in previous studies have had nodules larger than or equal to 1 cm (17,18). However, small thyroid nodules are prevalent in the general population (19-21), and there is controversy about whether thyroid nodules smaller than 1 cm should undergo biopsy before active surveillance (5,6,22,23).

A previous study by our team based on the same set of data has shown that the diagnostic efficacy of score-based RSSs is superior to that of pattern-based RSSs, but it did not classify and compare the diagnostic performance in detail according to the size of the nodules (22). Therefore, the assessment of the diagnostic performance should not be disregarded in small (<1 cm) thyroid nodules. This study set out to assess the diagnostic efficiency of the four score-based RSSs in relation to nodule size in the identification of thyroid cancer. We present this article in accordance with the STARD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-282/rc).

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Approval was granted by the Scientific Research And Clinical Trials Ethics Committee of The First Affiliated Hospital of Zhengzhou University of China (Date: 16 August 2022; No. 2022-KY-0974-001) and this study was granted a waiver of written informed consent for use of data by the Ethics Committee.

Study cohort

Data was collected from patients who underwent thyroid US examination at our institution, a tertiary referral center, between July 2013 and January 2019. A total of 2,744 patients with 4,075 thyroid nodules had undergone pathological diagnosis after thyroidectomy and/or US-guided FNA. Clinical decision makers took into account US imaging features, nodule size, patient age, underlying condition (such as symptoms, history of irradiation, cancer predisposition syndromes), and the patient's or parent's preference before performing a biopsy. Malignant nodules were defined as those that are histologically malignant after surgery or cytologically classified as Bethesda category VI. In total, 131 nodules in 77 patients were excluded from this study due to blurred US images, a lack of two vertical sections, or a lack of definitive cytopathologic results after FNA without surgical confirmation. Ultimately, 3,944 nodules in 2,667 patients (2,045 women and 622 men) were included in this study, of which 353 nodules underwent FNA and 3,591 nodules underwent thyroidectomy (*Figure 1*). The mean age of the patients was 47.2±12.2 years (ranging from 7 to 82 years), and the mean size of the 3,944 thyroid nodules was 16.9±14.5 mm (ranging from 1.5 to

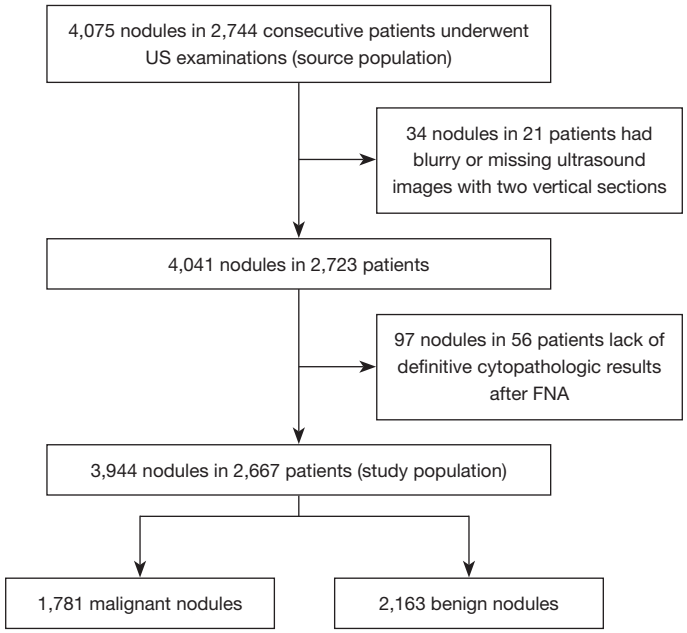

Figure 1 Flowchart showing the recruitment of study participants. US, ultrasound; FNA, fine-needle aspiration.

102.0 mm).

Evaluating and analyzing US examinations and imaging

A 5–14-MHz linear probe and a real-time US system (Aplio300; Toshiba, Tokyo, Japan) were used for all US examinations. The examinations were performed by a highly experienced senior radiologist (K.F.C.), with 34 years of expertise in thyroid imaging. All US examinations were performed in accordance with the American Institute of Ultrasound in Medicine (AIUM) (24) thyroid scanning protocol. In the US examination, each target nodule (thyroid nodules and suspicious cervical lymph nodes) typically receives at least one gray scale and one Doppler US image, covering both the transverse and longitudinal plane. Supplementary images were captured to highlight important US features in the US. The US data has been saved and stored on the internal hard disk for further offline analysis. The size of each nodule was determined by the maximal diameter of the US examination.

A senior radiologist (K.F.C.) with 34 years of experience in thyroid imaging led an overview and a discussion session to reach an agreement on the definitions of the US lexicons from the ACR-TIRADS and C-TIRADS (Table S1), which included size (the maximal diameter at US), echogenicity (hyperechoic, isoechoic, hypoechoic, markedly hypoechoic), composition (solid, predominately cystic, predominately solid, cystic, spongiform), orientation (vertical/taller-than-wide, horizontal/wider-than-tall), echogenic foci (punctate echogenic foci, peripheral calcifications, macrocalcification, comet-tail artifacts), and margins (smooth, irregular, lobulated, ill-defined, extrathyroidal extension). Subsequently, an interactive case-based training session was carried out by using 30 representative thyroid nodules that were not included in this study. In this study, the Kwak-TIRADS utilized the ACR's US lexicon of thyroid nodules, because it did not have its own US lexicon.

Two radiologists (C.F. and Y.J.H., with 14 and 13 years of clinical experience in thyroid US image evaluation and thyroid US scan performance, respectively) blinded to the biopsy results and the final pathological diseases independently reviewed the US features. If there was any disagreement, consensus would be used to draw conclusions. Without any prior knowledge of the FNA results or final diagnoses, Y.Y.C., the reviewer, classified the nodules using the assessed US features and determined each nodule's eligibility for FNA based on its size and category according to each RSS (Table S2).

Data and statistical analysis

In multiple previous studies, the diagnostic efficiency could

be computed based on the US categories (before applying size thresholds for biopsy) (9,25) and indications for FNA (after applying size thresholds for biopsy) (17,26). However, the diagnostic performance of the biopsy criteria have only been calculated in large thyroid nodules (≥10 mm).

Before applying size thresholds for FNA, the triages of the four RSSs were dichotomized into suggestive malignant nodules (category 4b to 5 for the Kwak-TIRADS and the C-TIRADS, category 4 to 5 for the ACR-TIRADS and the AI-TIRADS) and suggestive benign nodules (category 2 to 4a for the Kwak-TIRADS and the C-TIRADS, category 1 to 3 for the ACR-TIRADS and the AI-TIRADS) in accordance with the level of suspicion each category represents when calculating diagnostic performance. The dichotomy have been introduced in previous studies (14,27). After applying size thresholds for FNA, all nodules were dichotomized into those for which a biopsy was indicated (test positivity) and those for which it was not (test negativity).

The unnecessary FNA rate was calculated as the proportion of benign nodules recommended for FNA. The FNA rate was calculated as the proportion of the nodules recommended for FNA in all nodules. With sensitivity, specificity, accuracy, and AUCs, we evaluated the diagnostic performance.

The demographic data of benign and malignant nodules were compared using the independent two-sample *t*-test for numerical data (age and nodule size) and the chi-square or Fisher's exact test for categorical variables (sex and size distribution). The chi-square test was used to analyze sensitivity, specificity, accuracy, and unnecessary FNA rates among the four RSSs. When there was an overall difference between groups, the chi-square test was further used for pairwise comparisons, and the P value was adjusted by Bonferroni correction. The AUC and 95% confidence intervals (CIs) were generated and compared using the DeLong method or Z test. Statistical data were processed using the software SPSS 26.0 (IBM Corp., Armonk, NY, USA) and MedCalc version 18.2.1 (MedCalc Software, Mariakerke, Belgium). Two-sided P values <0.05 were regarded as indicative of statistical significance.

Results

Baseline clinicopathological characteristics

Out of the 3,944 thyroid nodules, 2,163 (54.8%) were benign whereas 1,781 (45.2%) were malignant. Of all the nodules, 43.5% (1,715 of 3,944) had a diameter less than 1 cm, and 56.5% (2,229 of 3,944) had a diameter of 1 cm or more. The average age of the benign group (49.3±12.1 years; range, 10–82 years) was greater than that of the malignant group (44.7±11.9 years); range, 7–82 years; P<0.001). At the same time, the nodules in the benign group were significantly larger than those in the malignant group (20.2±15.8 *vs.* 13.2±11.6 mm, P<0.001). Meanwhile, 901 male patients (22.8%) were significantly fewer than 3,043 female patients (77.2%), but there was no correlation between gender and the risk of malignant tumors (P=0.131). We summarized the demographics and US features of the patients and nodules [please refer to *Table 1* of our previous research (22)].

The most prevalent malignant nodules were papillary thyroid carcinomas [1,719 papillary thyroid carcinomas, including 47 follicular variant thyroid carcinomas, 18 medullary carcinomas, 22 follicular carcinomas, 4 lymphomas, 1 metastasis, 8 squamous cell carcinoma, 2 anaplastic carcinomas, 4 mixed carcinomas, 2 Hürthle cell carcinomas, and 1 poorly differentiated carcinoma (insular carcinoma)]. Nodular goiters were the most common benign nodules [1,696 nodular goiters, 37 follicular adenomas, 130 thyroiditis (including subacute, lymphocytic, and granulomatous), 12 Hürthle cell adenomas, 18 hemorrhagic cysts, 193 adenomatous goiter, 31 Graves' diseases, 39 simple goiters, 4 cysts, 1 cystic lymphangioma, and 2 neurilemmomas].

Malignancy rates and fraction of nodule counts by category

The detailed malignancy rates and fraction of nodule counts according to each RSS category is presented in *Figure 2* and Table S3. In our cohort, the overall malignancy rate was 45.2% (1,781 of 3,944), and significant differences in malignancy rates were observed among categories in each RSS (P<0.05 for all). Malignancy rates in each RSS increased across categories. Most categories had calculated malignancy risks within the suggested malignancy risk range for each RSS. The ACR-TIRADS and the AI-TIRADS had the highest proportion of nodule members in category 5 (TR5), whereas the C-TIRADS and the Kwak-TIRADS had the highest proportion in category 4c (TR4c).

Diagnostic performance in small nodules (<1 cm)

Table 2 presents the diagnostic performances of the four RSSs in nodules smaller than 1 cm. There was a slight

Table 1 Comparison of diagnostic performances between before and after applying size threshold for biopsy in large nodules

Systems	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC	FNA rate (%)	Unnecessary FNA rate (%)
ACR-TIRADS						
Before	95.5 (793/830) (94.1–96.9)	62.5 (875/1,399) (59.8–65.2)	74.8 (1,668/2,229) (73.1–76.6)	0.891 (0.877–0.904)	59.1 (1,317/2,229) (57.1–61.1)	39.8 (524/1317) (37.1–42.4)
After	90.0 (747/830) (88.0–92.0)	55.6 (778/1,399) (53.0–58.3)	68.4 (1,525/2,229) (66.5–70.3)	0.728 (0.709–0.746)	61.4 (1,368/2,229) (59.3–63.4)	45.4 (621/1,368) (42.9–48.2)
P value	<0.001	<0.001	<0.001	<0.001	0.009	0.003
AI-TIRADS						
Before	92.7 (769/830) (90.8–94.5)	71.6 (1,002/1,399) (69.3–73.9)	79.5 (1,771/2,229) (77.7–81.1)	0.904 (0.891–0.916)	52.3 (1,166/2,229) (50.2–54.3)	34.0 (397/1,166) (31.6–36.5)
After	90.8 (754/830) (88.8–92.8)	66.1 (925/1,399) (63.7–68.5)	75.3 (1,679/2,229) (73.5–77.2)	0.785 (0.767–0.802)	55.1 (1,228/2,229) (53.0–57.2)	38.6 (474/1,228) (36.1–41.3)
P value	0.032	<0.001	<0.001	<0.001	<0.001	0.021
Kwak-TIRADS						
Before	92.5 (768/830) (90.6–94.3)	71.3 (997/1,399) (68.8–73.6)	79.2 (1,765/2,229) (77.4–80.9)	0.904 (0.891–0.916)	52.5 (1,170/2,229) (50.4–54.5)	34.4 (402/1,170) (31.6–37.0)
After	97.8 (812/830) (97.0–98.8)	36.8 (515/1,399) (34.5–39.2)	59.5 (1,327/2,229) (57.5–61.7)	0.673 (0.653–0.693)	76.1 (1,696/2,229) (74.3–77.8)	52.1 (884/1,696) (49.6–54.6)
P value	<0.001	<0.001	<0.001	<0.0001	<0.001	<0.001
C-TIRADS						
Before	90.7 (753/830) (88.8–92.7)	71.8 (1,004/1,399) (69.3–74.2)	78.8 (1,757/2,229) (77.2–80.5)	0.885 (0.871–0.898)	51.5 (1,148/2,229) (49.5–53.5)	34.4 (395/1,148) (31.8–37.3)
After	88.7 (736/830) (86.5–90.7)	47.2 (660/1,399) (44.7–49.7)	62.6 (1,396/2,229) (60.6–64.7)	0.679 (0.659–0.699)	66.2 (1,475/2,229) (64.3–68.1)	50.1 (739/1,475) (47.4–52.5)
P value	0.118	<0.001	<0.001	<0.0001	<0.001	<0.001

Numbers in brackets are 95% confidence intervals. AUC, area under the curve; FNA, fine-needle aspiration; ACR, American College of Radiology; TIRADS, Thyroid Imaging Reporting and Data System; before, before applying size thresholds for biopsy; after, after applying size thresholds for biopsy; AI, artificial intelligence; C, Chinese.

difference in the diagnostic performance of the four RSSs. The accuracy was not significantly different among the four RSSs. The sensitivity was the highest in the ACR-TIRADS but it was not significantly different from that of the AI-TIRADS and the Kwak-TIRADS (97.7%, and 97.2%, 97.2%, respectively, P>0.05 for all), whereas it was the lowest with the C-TIRADS (94.2%, P<0.05 for all).

The specificity was the highest with the C-TIRADS and was not significantly different from that of the AI-TIRADS and the Kwak-TIRADS (55.0%, 49.9%, and 49.7%, respectively, P>0.05 for all). The Kwak-TIRADS had the highest AUC, which was not significantly different from the AI-TIRADS (0.841 versus 0.834, P=0.107).

Diagnostic performance in large nodules (≥1 cm)

Before the size thresholds for biopsy were applied (*Table 1*), there was a slight difference in the diagnostic performance of the four RSSs. The accuracy was the highest with the AI-TIRADS but was not significantly different from that of the C-TIRADS and the Kwak-TIRADS (79.5%, 78.8%, and 79.2%, respectively, P>0.05 for all). The sensitivity was the highest with the ACR-TIRADS but was not significantly different from that of the AI-TIRADS and the Kwak-TIRADS (95.5%, 92.7%, and 92.5%, respectively, P>0.05 for all). The specificity was the highest with the C-TIRADS but was not significantly different from that of the AI-

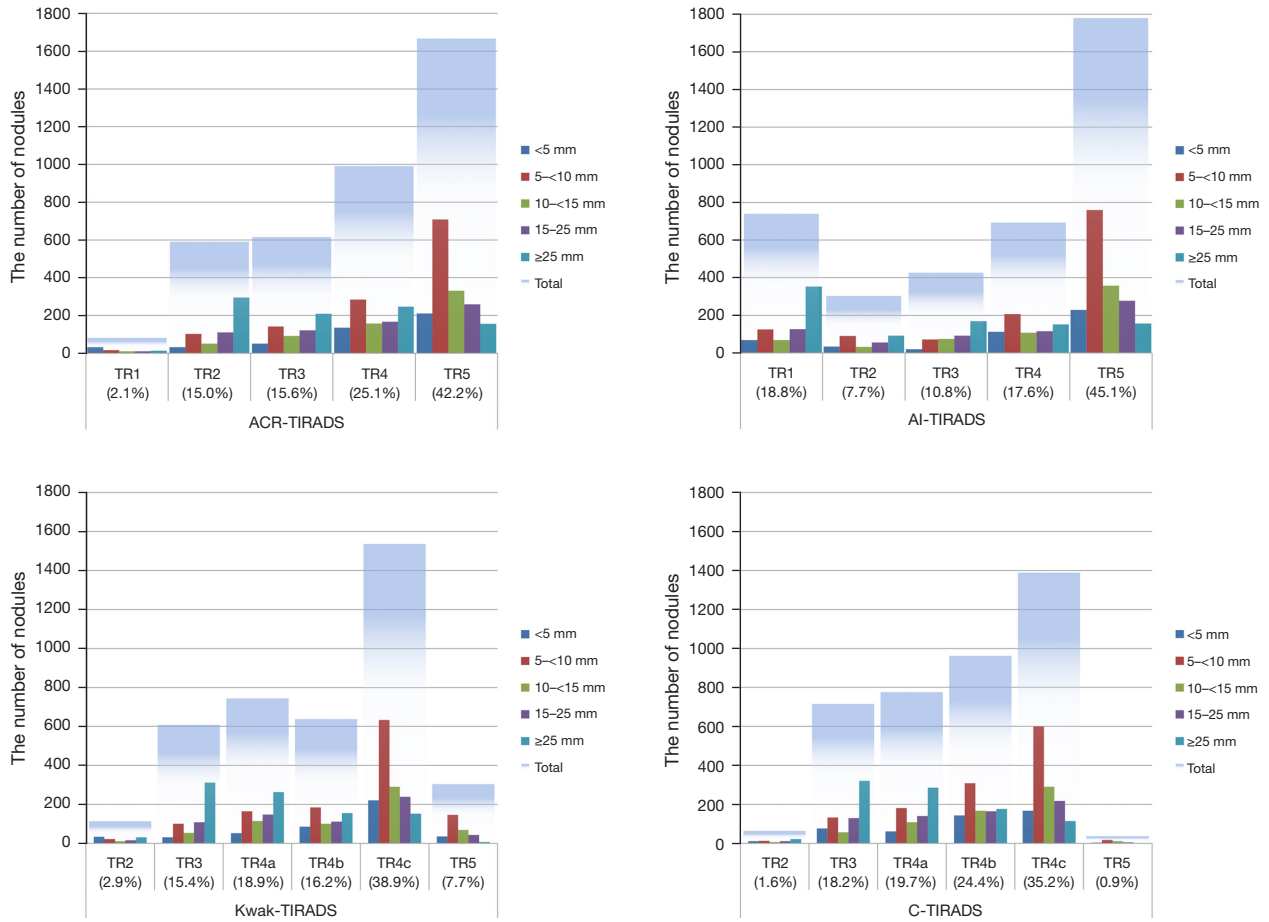

Figure 2 The proportion of nodule numbers and the size distribution of nodules in each category of the four RSSs. The percentages represent the proportion of nodule numbers in each category. TR, risk stratification category; RSS, risk stratification system; ACR, American College of Radiology; TIRADS, Thyroid Imaging Reporting and Data System; AI, artificial intelligence; C, Chinese.

Table 2 Diagnostic performances of the nodules <1 cm in the four RSSs

Systems	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC
ACR-TIRADS	97.7 (929/951) (96.7–98.6)	46.2 (353/764) (42.8–49.7)	74.8 (1,282/1,715) (72.7–76.7)	0.82 (0.801–0.838)
AI-TIRADS	97.2 (924/951) (96.0–98.2)	49.9 (381/764) (46.3–53.4)	76.1 (1,305/1,715) (74.1–78.0)	0.834 (0.816–0.851)
Kwak-TIRADS	97.2 (924/951) (96.1–98.2)	49.7 (380/764) (46.1–53.3)	76.0 (1,304/1,715) (74.1–78.1)	0.841 (0.823–0.858)
C-TIRADS	94.2 (896/951) (92.6–95.7)	55.0 (420/764) (51.3–58.5)	76.7 (1,316/1,715) (74.7–78.8)	0.811 (0.791–0.829)

Numbers in brackets are 95% confidence intervals. RSS, risk stratification system; AUC, area under the curve; ACR, American College of Radiology; TIRADS, Thyroid Imaging Reporting and Data System; AI, artificial intelligence; C-TIRADS, Chinese Thyroid Imaging, Reporting and Data System.

Table 3 Comparison of diagnostic performances between small and large nodules before applying size thresholds for biopsy

Systems	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC
ACR-TIRADS				
<10 mm	97.7	46.2	74.8	0.820
≥10 mm	95.5	62.5	74.8	0.891
χ^2/Z value	6.363	53.764	0.003	6.021
P value	0.012	<0.001	0.954	<0.05
AI-TIRADS				
<10 mm	97.2	49.9	76.1	0.834
≥10 mm	92.7	71.6	79.5	0.904
χ^2/Z value	19.195	101.416	6.373	6.137
P value	<0.001	<0.001	0.012	<0.05
Kwak-TIRADS				
<10 mm	97.2	49.7	76	0.841
≥10 mm	92.5	71.3	79.2	0.904
χ^2/Z value	20.019	98.987	5.566	5.582
P value	<0.001	<0.001	0.018	<0.05
C-TIRADS				
<10 mm	94.2	55.0	76.7	0.811
≥10 mm	90.7	71.8	78.8	0.885
χ^2/Z value	7.883	61.945	2.46	6.027
P value	0.005	<0.001	0.117	<0.05

AUC, area under the curve; ACR, American College of Radiology; TIRADS, Thyroid Imaging Reporting and Data System; AI, artificial intelligence; C, Chinese.

TIRADS and the Kwak-TIRADS (71.8%, 71.6%, and 71.3%, respectively, P>0.05 for all). The Kwak-TIRADS had the highest AUC (0.904), which was not significantly different from that of the AI-TIRADS (0.904, P=0.86). The AI-TIRADS had the lower FNA rate and unnecessary FNA rate (52.3%, 34.0%), which were similar to those of the Kwak-TIRADS (52.5% and 34.4%, respectively) and C-TIRADS (51.5% and 34.4%, respectively, P>0.05 for all).

After applying size thresholds for biopsy (*Table 1*), the sensitivity was the highest with the Kwak-TIRADS (97.8%) and was not significantly different between the ACR-TIRADS, AI-TIRADS, and C-TIRADS (90.0%, 90.8%, and 88.7%, respectively, P>0.05 for all). The specificity and accuracy were highest with the AI-TIRADS (66.1% and 75.3%, respectively) and the lowest with the Kwak-TIRADS (36.8% and 59.5%, respectively), whereas the accuracy was similar between Kwak-TIRADS and C-TIRADS (59.5% *vs.* 62.6%, respectively, P>0.05). The AUC was highest in the AI-TIRADS (0.785), followed by the ACR-TIRADS (0.728) and the lowest with the C-TIRADS and Kwak-TIRADS (0.679 and 0.673, respectively, P=0.455). The lowest FNA rate and unnecessary FNA rate was observed in the AI-TIRADS (55.1% and 38.6%, respectively, P<0.05 for all).

Comparison of diagnostic performances between before and after applying size thresholds for biopsy in large nodules

Table 1 summarizes the diagnostic performances of the four RSSs in large nodules, both before and after applying the size threshold of biopsy. Notably, the diagnostic performances of the four RSSs before applying size thresholds of the biopsy were more efficient than they were after, except for the sensitivity in the C-TIRADS (90.7% *vs.* 88.7%, P=0.118). However, the FNA rate and unnecessary FNA rate of the RSSs before applying size thresholds of biopsy were lower than they were after (ACR-TIRADS, 59.1% *vs.* 61.4%, 39.8% *vs.* 45.4%; AI-TIRADS, 52.3% *vs.* 55.1%, 34.0% *vs.* 38.6%; Kwak-TIRADS, 52.5% *vs.* 76.1%, 34.4% *vs.* 52.1%; C-TIRADS, 51.5% *vs.* 66.2%, 34.4% *vs.* 50.1%, P<0.05 for all).

Comparison of diagnostic performances between small and large nodules before applying size thresholds for biopsy

The four RSS showed comparable diagnostic performance for small (<1 cm) and large (≥1 cm) nodules before applying size thresholds for biopsy (in large nodules, sensitivity, 90.7–95.5%; specificity, 62.5–71.8%; accuracy, 74.8–79.5%; AUC, 0.885–0.904. In small nodules, sensitivity, 94.2–97.7%; specificity, 46.2–55.0%; accuracy, 74.8–76.7%; AUC, 0.811–0.841). The small nodules showed higher sensitivity and lower specificity than the large nodules (ACR-TIRADS, 97.7% *vs.* 95.5%, 46.2% *vs.* 62.5%; AI-TIRADS, 97.2% *vs.* 92.7%, 49.9% *vs.* 71.6%; Kwak-TIRADS, 97.2% *vs.* 92.5%, 49.7% *vs.* 71.3%; C-TIRADS, 94.2% *vs.* 90.7%, 55.0% *vs.* 71.8%, respectively, all P<0.05). The specific results are shown in *Table 3*.

Discussion

This study aimed to compare the diagnostic performance of four current score-based RSSs in diagnosing small (<1 cm)

and large (≥1 cm) nodules. The diagnostic performances of the RSSs before applying a biopsy size threshold were more efficient than they were after it, except for the sensitivity of C-TIRADS (90.7% *vs.* 88.7%, P=0.118) in large nodules. The four RSSs showed effective diagnostic performance in both large and small nodules, with small nodules having higher sensitivity and lower specificity than the large nodules.

Various scientific societies and individuals have proposed the thyroid imaging RSSs as an initial method to stratify malignancy risk (3-6,28). Although many RSSs have US triages that overlap with similar estimated risks of malignancy, the size threshold for recommending FNA and the accuracy differ among each RSS (29,30). Although a 10 mm size threshold is commonly used to indicate biopsy for the highest triage, the size threshold for other triages varies (4,6,15). Previous studies have indicated that disparities in diagnostic performance among RSSs are primarily due to varying biopsy size thresholds, with simulated size thresholds for FNA (17,31). Simulation studies have shown that the diagnostic performance of the RSSs is similar at the same size threshold for biopsy (31,32). Additionally, the diagnostic performance estimated through the US categories was comparable among each RSS (9). This is also similar to a previous study by our team wherein we examined the diagnostic performance and unnecessary FNA rates of six RSSs using the same size thresholds provided by ACR-TIRADS and US based final assessment categories, respectively (22).

Our findings suggested that, in the current score-based RSSs, managing large nodules using only US categories would be an effective strategy, which is equal to lowering size thresholds (same size thresholds, ≥1 cm) for suggestive malignant nodules. Before applying size thresholds for biopsy, the diagnostic performance of the four RSSs varied slightly in large nodules, with AUCs ranging from 0.885 to 0.904 and accuracy ranging from 74.8% to 79.5%. Furthermore, the diagnostic performances of the four RSSs before applying size threshold of biopsy were more efficient than that after it, except for the sensitivity of C-TIRADS (90.7% *vs.* 88.7%, P=0.118) in large nodules. Kim *et al.* concluded that permitting biopsy of the highest triage nodules smaller than 1 cm in the pediatric population improved the diagnostic performances of the biopsy criteria of five RSSs (26). In line with this, the 2021 Korean Thyroid Imaging Reporting and Data System (K-TIRADS) dropped the biopsy threshold for K-TIRADS 4 and 5 triage and showed superior diagnostic accuracy to the 2016 K-TIRADS in pediatric patients (33). By employing smaller size thresholds for biopsy, we also showed in our earlier study that the diagnostic performances of the C-TIRADS and Kwak-TIRADS were more effective (34). However, studies have supposed that since ACR-TIRADS has relatively larger biopsy size thresholds for nodules categorized as moderately and mildly suspicious category (17,32), it contributes to a higher level of specificity and a reduced rate of unnecessary FNA (35-37). All enrolled nodules had undergone FNA or surgery in the past, even if this indication may not be based on RSSs (38,39). In a study by Dong *et al.* (40), by comparing the consistency of ACR and TIRADS recommendations, it was shown that the two guidelines have good consistency for thyroid nodules ≥1.5 cm. However, the article points out that the diagnostic efficacy of ACR is higher than that of C-TIRADS, which is slightly different from this study. It is considered that this difference is due to the different proportion of nodules and the distribution of benign and malignant nodules in each category in the recruited sample.

This study is the first to compare the diagnostic effectiveness of four current score-based RSSs based on thyroid nodule size. Despite the fact that FNA is now frequently applied by nodules larger than 1 cm (4,6,15), the diagnosis of nodules less than 1 cm in size should not be disregarded. There is controversy about whether thyroid nodules smaller than 1 cm should undergo biopsy before active surveillance (5,6,23). The four RSSs showed efficient diagnostic performance in this investigation in both large and small nodules. In comparison to large nodules, small nodules showed higher sensitivity and lower specificity. In small nodules, the highest AUC was observed with the Kwak-TIRADS and without significant differences to that of the AI-TIRADS (0.841 *vs.* 0.834, P=0.107). The specificity of the C-TIRADS was the highest, but was not significantly different to that of the AI-TIRADS and the Kwak-TIRADS (55.0%, 49.9%, and 49.7%, respectively, P>0.05 for all).

These findings mean that the role of US RSS in comparison to size thresholds is an important consideration. The diagnostic performance of the RSSs is efficient in both large and small nodules, indicating that although size thresholds play a role, the specific characteristics defined by the US categories can provide valuable information for diagnosis and management decisions. This might prompt clinicians to reevaluate the balance between relying on size thresholds and paying more attention to the detailed US features and categories when dealing with thyroid

nodules. Furthermore, the differences in the performance of the various RSSs highlight the need for individualized approaches based on the specific characteristics and context of each patient. Clinicians may need to carefully consider which RSS is most appropriate for a given situation, taking into account factors such as the patient's age, overall health, and the specific features of the nodules. This could lead to more tailored and effective management strategies for thyroid nodules, especially for those that are less than 1 cm in size. Overall, these findings call for a more nuanced understanding and application of RSSs in the clinical management of thyroid nodules.

However, several limitations remain in this study that need to be addressed in the future. Firstly, only the patients who underwent surgery or FNA in a tertiary referral center were included in this series, resulting in a higher proportion of malignant nodules (45.2%) and potential bias. This may not accurately reflect the true prevalence of the general population, which may overestimate the risk of malignant tumors, leading to more proactive management recommendations that may not be applicable to all situations. Secondly, the inclusion of a significant number of cases verified by surgical pathology may have resulted in a few false positives and false negatives. Surgeons can label specimens for pathologists (one of the largest or most suspicious thyroid nodules per thyroid lobe) to help reduce this falsehood. Thirdly, the absence of interobserver agreement assessment for US features of the thyroid nodules may have impacted image homogeneity. In future research, we can include radiologists with different US work experiences for future consistency evaluation to determine whether this method is suitable for radiologists with different experiences. Finally, when calculating diagnostic performance according to US categories, the cutoffs for each RSS may influence the results. Previous studies have shown that different puncture thresholds can lead to different diagnostic results. In the future, we can set different thresholds for further research in this study area. The possibility of generalizing research results is limited, as there may have been selection bias from specific study populations. Future research may include more diverse patients from different medical environments and geographical locations.

Conclusions

A potential effective strategy for managing large nodules in the current score-based RSSs could be to rely solely on US categories rather than size thresholds for biopsy. Before applying size thresholds for biopsy, the four RSSs showed effective diagnostic performance in both large and small nodules, with small nodules (<1 cm) exhibiting higher sensitivity and lower specificity than large nodules. Our findings provided a novel management approach for large nodules and provide a basis for managing small nodules. Nevertheless, further validation of our results through larger multicenter studies will be crucial to enhance individual management of thyroid nodules.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-282/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-282/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Approval was granted by the Scientific Research and Clinical Trials Ethics Committee of The First Affiliated Hospital of Zhengzhou University of China (Date: 16 August 2022; No. 2022-KY-0974-001) and this study was granted a waiver of written informed consent for use of data by the Ethics Committee.

References

1. Ha EJ, Lim HK, Yoon JH, Baek JH, Do KH, Choi M, Choi JA, Lee M, Na DG; Korean Society of Thyroid Radiology (KSThR) and Korean Society of Radiology. Primary Imaging Test and Appropriate Biopsy Methods for Thyroid Nodules: Guidelines by Korean Society of Radiology and National Evidence-Based Healthcare Collaborating Agency. Korean J Radiol 2018;19:623-31.
2. Zhou J, Yin L, Wei X, Zhang S, Song Y, Luo B, et al. 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. Endocrine 2020;70:256-79.
3. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur Thyroid J 2017;6:225-37.
4. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, Stavros AT. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14:587-95.
5. Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al. Ultrasonography Diagnosis and Imaging-Based Management of Thyroid Nodules: Revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol 2016;17:370-95.
6. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1-133.
7. Yoon SJ, Na DG, Gwon HY, Paik W, Kim WJ, Song JS, Shim MS. Similarities and Differences Between Thyroid Imaging Reporting and Data Systems. AJR Am J Roentgenol 2019;213:W76-84.
8. Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH. Unnecessary thyroid nodule biopsy rates under four ultrasound risk stratification systems: a systematic review and meta-analysis. Eur Radiol 2021;31:2877-85.
9. Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH. Diagnostic Performance of Four Ultrasound Risk Stratification Systems: A Systematic Review and Meta-Analysis. Thyroid 2020;30:1159-68.
10. Barbaro D, Simi U, Meucci G, Lapi P, Orsini P, Pasquini C. Thyroid papillary cancers: microcarcinoma and carcinoma, incidental cancers and non-incidental cancers - are they different diseases? Clin Endocrinol (Oxf) 2005;63:577-81.
11. Gao L, Xi X, Jiang Y, Yang X, Wang Y, Zhu S, Lai X, Zhang X, Zhao R, Zhang B. Comparison among TIRADS (ACR TI-RADS and KWAK- TI-RADS) and 2015 ATA Guidelines in the diagnostic efficiency of thyroid nodules. Endocrine 2019;64:90-6.
12. Schenke S, Zimny M. Combination of Sonoelastography and TIRADS for the Diagnostic Assessment of Thyroid Nodules. Ultrasound Med Biol 2018;44:575-83.
13. Migda B, Migda M, Migda MS, Slapa RZ. Use of the Kwak Thyroid Image Reporting and Data System (K-TIRADS) in differential diagnosis of thyroid nodules: systematic review and meta-analysis. Eur Radiol 2018;28:2380-8.
14. 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;30:3793-802.
15. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, Jung HK, Choi JS, Kim BM, Kim EK. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. Radiology 2011;260:892-9.
16. Wildman-Tobriner B, Buda M, Hoang JK, Middleton WD, Thayer D, Short RG, Tessler FN, Mazurowski MA. Using Artificial Intelligence to Revise ACR TI-RADS Risk Stratification of Thyroid Nodules: Diagnostic Accuracy and Utility. Radiology 2019;292:112-9.
17. Huh S, Yoon JH, Lee HS, Moon HJ, Park VY, Kwak JY. Comparison of diagnostic performance of the ACR and Kwak TIRADS applying the ACR TIRADS' size thresholds for FNA. Eur Radiol 2021;31:5243-50.
18. 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.
19. Castro MR, Gharib H. Continuing controversies in the management of thyroid nodules. Ann Intern Med 2005;142:926-31.
20. Davies L, Morris LG, Haymart M, Chen AY, Goldenberg

D, Morris J, Ogilvie JB, Terris DJ, Netterville J, Wong RJ, Randolph G; AACE Endocrine Surgery Scientific Committee. AMERICAN ASSOCIATION OF CLINICAL ENDOCRINOLOGISTS AND AMERICAN COLLEGE OF ENDOCRINOLOGY DISEASE STATE CLINICAL REVIEW: THE INCREASING INCIDENCE OF THYROID CANCER. Endocr Pract 2015;21:686-96.
21. Hegedüs L. Clinical practice. The thyroid nodule. N Engl J Med 2004;351:1764-71.
22. Fu C, Cui Y, Li J, Yu J, Wang Y, Si C, Cui K. Effect of the categorization method on the diagnostic performance of ultrasound risk stratification systems for thyroid nodules. Front Oncol 2023;13:1073891.
23. Brito JP, Ito Y, Miyauchi A, Tuttle RM. A Clinical Framework to Facilitate Risk Stratification When Considering an Active Surveillance Alternative to Immediate Biopsy and Surgery in Papillary Microcarcinoma. Thyroid 2016;26:144-9.
24. AIUM Practice Parameter for the Performance of a Thyroid and Parathyroid Ultrasound Examination. J Ultrasound Med 2016;35:1-11.
25. Hekimsoy İ, Öztürk E, Ertan Y, Orman MN, Kavukçu G, Özgen AG, Özdemir M, Özbek SS. Diagnostic performance rates of the ACR-TIRADS and EU-TIRADS based on histopathological evidence. Diagn Interv Radiol 2021;27:511-8.
26. Kim PH, Yoon HM, Baek JH, Chung SR, Choi YJ, Lee JH, Lee JS, Jung AY, Cho YA, Bak B, Na DG. Diagnostic Performance of Five Adult-based US Risk Stratification Systems in Pediatric Thyroid Nodules. Radiology 2022;305:190-8.
27. Kim TY, Shong YK. Active Surveillance of Papillary Thyroid Microcarcinoma: A Mini-Review from Korea. Endocrinol Metab (Seoul) 2017;32:399-406.
28. Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L, Paschke R, Valcavi R, Vitti P. AMERICAN ASSOCIATION OF CLINICAL ENDOCRINOLOGISTS, AMERICAN COLLEGE OF ENDOCRINOLOGY, AND ASSOCIAZIONE MEDICI ENDOCRINOLOGI MEDICAL GUIDELINES FOR CLINICAL PRACTICE FOR THE DIAGNOSIS AND MANAGEMENT OF THYROID NODULES--2016 UPDATE. Endocr Pract 2016;22:622-39.
29. 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;283:560-9.
30. Hoang JK, Middleton WD, Farjat AE, Langer JE, Reading CC, Teefey SA, Abinanti N, Boschini FJ, Bronner AJ, Dahiya N, Hertzberg BS, Newman JR, Scanga D, Vogler RC, Tessler FN. Reduction in Thyroid Nodule Biopsies and Improved Accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. Radiology 2018;287:185-93.
31. Yim Y, Na DG, Ha EJ, Baek JH, Sung JY, Kim JH, Moon WJ. Concordance of Three International Guidelines for Thyroid Nodules Classified by Ultrasonography and Diagnostic Performance of Biopsy Criteria. Korean J Radiol 2020;21:108-16.
32. Ha SM, Baek JH, Na DG, Suh CH, Chung SR, Choi YJ, Lee JH. Diagnostic Performance of Practice Guidelines for Thyroid Nodules: Thyroid Nodule Size versus Biopsy Rates. Radiology 2019;291:92-9.
33. Kim PH, Yoon HM, Baek JH, Chung SR, Choi YJ, Lee JH, Lee JS, Jung AY, Cho YA, Bak B, Na DG. Diagnostic performance of the 2021 Korean thyroid imaging reporting and data system in pediatric thyroid nodules. Eur Radiol 2023;33:172-80.
34. Fu C, Cui Y, Li J, Wang Y, Si C, Cui K. The feasibility of decreasing the thresholds for biopsy in Kwak and C TIRADSs. Front Oncol 2023;13:1027802.
35. Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, Desser TS. Comparison of Performance Characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association Guidelines. AJR Am J Roentgenol 2018;210:1148-54.
36. Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, Maranghi M, Falcone R, Ramundo V, Cantisani V, Filetti S, Durante C. Reducing the Number of Unnecessary Thyroid Biopsies While Improving Diagnostic Accuracy: Toward the "Right" TIRADS. J Clin Endocrinol Metab 2019;104:95-102.
37. Ruan JL, Yang HY, Liu RB, Liang M, Han P, Xu XL, Luo BM. 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;29:4871-8.
38. Qi Q, Zhou A, Guo S, Huang X, Chen S, Li Y, Xu P. 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.
39. Castellana M, Castellana C, Treglia G, Giorgino F,

Giovanella L, Russ G, Trimboli P. Performance of Five Ultrasound Risk Stratification Systems in Selecting Thyroid Nodules for FNA. J Clin Endocrinol Metab 2020;105:dgz170.

40. Dong W, Wu Y, Cai T, Wang X. Comparison of diagnostic performance and FNA management of the ACR-TIRADS and Chinese-TIRADS based on surgical histological evidence. Quant Imaging Med Surg 2023;13:1711-22.

Cite this article as: Si CF, Yu J, Cui YY, Huang YJ, Cui KF, Fu C. Comparison of diagnostic performance of the current score-based ultrasound risk stratification systems according to thyroid nodule size. Quant Imaging Med Surg 2024;14(12):9234-9245. doi: 10.21037/qims-24-282