




# A new discriminant strategy combined with four TIRADS screening procedures increases ultrasound diagnostic accuracy—focusing on “wrong diagnostic” thyroid nodules

Ke Lu<sup>1,2</sup> · Long Wang<sup>2</sup> · Shuiqing Lai<sup>2</sup> · Zhijiang Chen<sup>2</sup> · Shuzhen Cong<sup>4</sup> · Chunwang Huang<sup>4</sup> · Kehong Gan<sup>4</sup> · Haixia Guan<sup>2</sup> · Jian Kuang<sup>2,3</sup> 

Received: 26 April 2022 / Revised: 10 August 2022 / Accepted: 19 August 2022 / Published online: 28 September 2022  
© The Author(s) 2022

## Abstract

**Objective** To utilize the discrepancies of different TIRADS, including ACR-TIRADS, Kwak-TIRADS, C-TIRADS, and EU-TIRADS, to explore methods for improving ultrasound diagnostic accuracy.

**Methods** In total, 795 nodules with cytological or surgical pathology were included. All nodules were screened by the four TIRADS according to their diagnostic concordance (Screening procedures, SP). Discriminant strategy (DS) derived from predictor variables was combined with SP to construct the evaluation method (SP+DS). The diagnostic performance of the SP+DS method alone and its derivational methods and two-TIRADS combined tests was evaluated.

**Results** A total of 86.8% (269/310) malignant nodules and 93.6% (365/390) benign cases diagnosed by the four TIRADS simultaneously were pathologically confirmed, while 12.0% (95/795) nodules could not be consistently diagnosed by them. The criteria of DS were that iso- or hyper-echogenicity nodules should be considered benign, while hypo- or marked hypo-echogenicity nodules malignant. For 95 inconsistently diagnosed nodules screened by at least two TIRADS, DS performed best with an accuracy of 79.0%, followed by Kwak-TIRADS (72.6%). In the overall sample, the sensitivity and AUC were highest for the SP+DS method compared to the four TIRADS (91.3%, 0.895). Combining ACR-TIRADS and Kwak-TIRADS via parallel test resulted in significant improvements in the sensitivity and AUC compared to ACR-TIRADS (89.2% vs. 81.4%, 0.889 vs. 0.863). Combining C-TIRADS and DS in serial resulted in the highest AUC (0.887), followed by Kwak-TIRADS (0.884), while EU-TIRADS was the lowest (0.879).

**Conclusions** For undetermined or suspected thyroid nodules, two-TIRADS combined tests can be used to improve diagnostic accuracy. Otherwise, considering the inconsistent diagnosis of two TIRADS may require attention to the echo characteristics to differentiate between benign and malignant nodules.

## Key Points

- *The discrepancies in the diagnostic performance of different TIRADS arise from their performance on inconsistently diagnosed nodules.*
- *ACR-TIRADS improves sensitivity via combining with Kwak-TIRADS in parallel (from 81.4 to 89.2%), while C-TIRADS increases specificity via combining with EU-TIRADS in serial (from 80.9 to 85.7%).*
- *If the diagnostic findings of two TIRADS are inconsistent, echo characteristics will be helpful for the differentiation of benign and malignant nodules with an accuracy of 79.0%.*

---

Ke Lu, Long Wang, and Shuiqing Lai have equal contributions.

---

Haixia Guan and Jian Kuang have equal contributions.

✉ Haixia Guan  
hxguan@vip.126.com

✉ Jian Kuang  
kuangjian@gdph.org.cn

<sup>1</sup> Department of Endocrinology, First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China

<sup>2</sup> Department of Endocrinology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong, China

<sup>3</sup> The Second School of Clinical Medicine, Southern Medical University, Guangzhou, Guangdong, China

<sup>4</sup> Department of Ultrasound, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong, China

**Keywords** Thyroid nodules · Ultrasound · TIRADS · Accuracy

### Abbreviations

ACR-TIRADS	American College of Radiology Thyroid Imaging Reporting and Data System
C-TIRADS 2020	Chinese Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules
DS	Discriminant strategy
EU-TIRADS	European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults
FNA	Fine-needle aspiration
Kwak-TIRADS	Kwak Thyroid Imaging Reporting and Data System
NPV	Negative predictive value
PPV	Positive predictive value
PTC	Papillary thyroid carcinoma
SP	Screening procedures
SP+A/C/K/E+DS	The evaluation method consists of the four TIRADS screening procedures with partially inconsistently diagnosed nodules judged by ACR-TIRADS/C- TIRADS/Kwak-TIRADS/EU- TIRADS and combined with discriminant strategy
SP+DS	The evaluation method consists of the four TIRADS screening procedures with partially inconsistently diagnosed nodules judged by discriminant strategy

### Introduction

Ultrasonography, as a simple, non-invasive diagnostic method, now occupies a priority position in the thyroid nodule evaluation process [1]. Certain ultrasound indices are significantly associated with thyroid cancer. Commonly used real-time ultrasound indices include size, composition, shape, halo sign, echogenicity, calcification, and some accessory features, including extrathyroidal extension, lymph nodes, blood flow, and elasticity. In fact, the sensitivity and specificity of any single ultrasound feature for diagnosing thyroid cancer are difficult to reach more than 90% simultaneously. Hypoechoic and solid nodules have higher diagnostic sensitivity but lower specificity, while nodules with microcalcifications, infiltrative margins, and taller-than-wide shapes have higher specificity but lower sensitivity [2].

Therefore, an ultrasound model consisting of a combination of valid ultrasound features is more helpful for identifying the nature of nodules.

In 2009, Chilean scholars first introduced the concept of TIRADS and defined ten ultrasound patterns to distinguish benign and malignant thyroid nodules [3]. Kwak then proposed a simplified stratified assessment system containing only five ultrasound indices including shape, echogenicity, structure, calcification, and margin in 2011 [4]. Subsequently published TIRADS, including ATA guidelines, EU-TIRADS, ACR-TIRADS, and KTA/KSThR-TIRADS, also have been constructed based on these five ultrasound modes. These TIRADS are currently clinically validated and have good diagnostic value. But the definitions of some features within their ultrasound lexicons (e.g., hypoechoic, solid, spongiform) are currently not uniform. And the number of assessment classifications, specific malignant features involved, and even the ways in utilizing suspicious ultrasonic features vary (i.e., calculating the number of suspicious features or using ultrasonic pattern for risk stratification), which make the malignancy rate of the classification from low suspicion to high suspicion different among these systems [5–8].

There is no perfect TIRADS to date. Various TIRADS have their own advantages. For example, the EU-TIRADS and ATA guidelines are pattern-dependent systems characterized by a high negative predictive value and sensitivity, whereas the ACR-TIRADS is a typical score-based system with a high positive predictive value and specificity [9–11]. We assume that various TIRADS will probably form complementary relationships based on these facts. For example, some TIRADS are more applicable in some thyroid nodular cases, while other TIRADS cannot classify them correctly. Further, is it possible to explore new methods to improve the diagnostic accuracy based on data from those unmatched nodules?

Thus, in this study, we focused on the differences of unmatched findings among four TIRADS (including the newly released C-TIRADS) [12]. We then explored potential ways such as two-TIRADS parallel or serial tests or one TIRADS combined with specific ultrasound features to improve the diagnostic accuracy.

### Methods

This retrospective study was approved by the Institutional Review Board, and the requirement for informed consent to review images and medical records was waived.

## Patients

From February 2016 to February 2019, 1001 thyroid nodules in 933 patients were enrolled in the study. Only definitely diagnosed nodules were included, malignant nodules were confirmed by surgical pathology, and benign nodules were diagnosed by surgical pathology or repeated Bethesda II findings. Based on the above criteria, 795 nodules were finally included, which involved 334 surgical malignant nodules, 63 benign surgical nodules and 398 nodules with repeated Bethesda II results. One hundred eighty-eight nodules that could not be clearly diagnosed were excluded: 7 nodules with Bethesda I cytopathology and 28 Bethesda III-V nodules with no further surgical pathology, 132 nodules with a single benign cytopathologic result, and 21 nodules with initial benign pathology but with increased nodular size in the follow-up period by ultrasound examination (mean interval of 21 months, range 2 to 35 months) (Fig. 1).

## Sonography examination and image evaluation

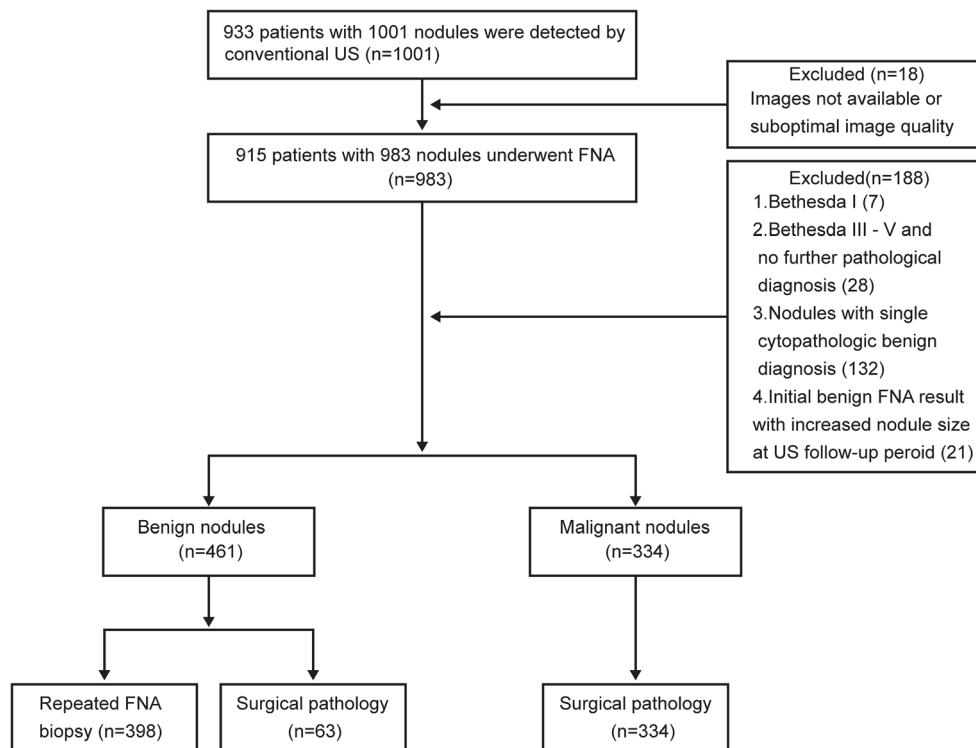
Conventional ultrasound examinations were performed using Aplio 500 (Toshiba Medical System), HI Vision Ascendus (Hitachi Medical Corporation), or HI Vision Preirus (Hitachi Medical Corporation) ultrasound instruments equipped with 5–12-MHz linear array transducers by board-certified radiologists. The ultrasound images were reviewed by one radiologist with more than 20 years of experience in thyroid ultrasound diagnosis and recorded by two experienced endocrinologists with the

help of the radiologist. They were all blinded to the patients' fine-needle aspiration (FNA) results or pathological diagnosis before sonography examination. In case of disagreement, conclusions would be drawn by consensus. Before assessing nodules, we studied and compared the lexicon and classification of four TIRADS (Supplementary Tables 1 and 2). The definition and classification of the various TIRADS regarding composition, echogenicity, margin, shape, and calcification are substantially similar. However, there are slight differences in the definition of solid, spongiform, hypoechoic, and section to evaluate the nodular orientation. C-TIRADS and Kwak-TIRADS are both counting-based systems. C-TIRADS included marked hypo-echogenicity and ill-defined margin into the scoring system and considered the presence of comet tail artifacts as a minus item.

## FNA, cytopathology, and histopathology

Thyroid nodules were judged as benign or malignant according to FNA cytopathology or surgical histopathology. The surgical pathological diagnosis was based on the WHO diagnostic criteria [13], and the cytopathological classification was based on the Bethesda system of thyroid FNA cytology proposed by the National Cancer Institute [14]. Informed consent was obtained from all patients before the FNA biopsy. The procedure was performed by an endocrinologist experienced in puncture using ultrasound-guided FNA technique by a color doppler ultrasound scanner with an L14-5 high-frequency line array probe (Ultrasonix Medical Ltd., Sonix SP). Benign

Fig. 1 Study flowchart



pathology was defined by repeated Bethesda II results according to the recommendation by the guidelines about ablation treatment [1, 15]. At our institution, the requirement of repeated FNAs meets the following situations: (1) The puncture results are Bethesda I, III, and IV, requiring repeated confirmation or performing further genetic test; (2) Nodules are categorized as intermediate or high suspicion according to TIRADS assessment, but the puncture results are Bethesda II; (3) Patients are scheduled to undergo thermal ablation treatment; (4) During the follow-up period, nodules are with a rapid increase in diameter or volume, or development of new suspicious features including margin, echogenicity, calcification, etc. The time interval was 2–4 weeks between two repeated benign FNAs.

### Statistical analysis

SPSS 26 software (IBM) and MedCalc 19.0.4 software (MedCalc) were used for statistical analysis. Quantitative data conforming to normal distribution were presented as mean  $\pm$  standard deviation and evaluated by independent samples t-test. Measurement data that did not conform to a normal distribution were expressed as median and interquartile ranges and evaluated by a nonparametric test. Qualitative data were expressed as frequencies and evaluated by a chi-squared test. The optimal cut-off point of each TIRADS was determined from ROC analysis when the Youden index was the highest, as well as sensitivity and specificity. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) were calculated. The McNemar test was used to assess the differences in parameters. Stepwise discriminant analysis was done to determine variables that may discriminate between benign and malignant nodules. Two-sided  $p$  values  $< 0.05$  were considered statistically significant.

## Results

### Baseline

There were no significant differences in the age and gender of patients with benign and malignant nodules. The maximum size of malignant nodules was significantly smaller than that of benign nodules (median 1.00 [Q1–Q3, 0.80–1.50] vs. 2.40 [Q1–Q3, 1.50–3.20]) (Supplementary Table 3). The number of surgical pathological papillary thyroid carcinoma (PTC) was 334. There were 52 surgical pathologies of nodular goiter, three adenomatous goiters, six follicular thyroid adenoma, and two nodular Hashimoto thyroiditis among benign nodules.

As shown in Table 1, the malignant rate of ACR-TIRADS TR3 and TR4, C-TIRADS CTR 4b, and EU-TIRADS grade 4 were higher than the recommended malignancy rate. There

were significant differences between the four TIRADS grades (all  $p < 0.001$ ).

Supplementary Table 4 shows the diagnostic performance of the four TIRADS. The results showed that the best diagnostic cut-off values of ACR-TRADS, Kwak-TIRADS, C-TIRADS and EU-TIRADS were TR5, 4c, CTR 4b, and grade 5, respectively. C-TIRADS had the highest sensitivity (91.6%) and NPV (93.0%), while ACR-TIRADS had the highest specificity (91.1%) and PPV (86.9%). However, Kwak-TIRADS had the highest AUC 0.884 (95% CI:0.860–0.906).

Figure 2 illustrates the diagnostic distribution for the four TIRADS in assessing pathological benign and malignant nodules. The number of cases with inconsistent findings of benign pathology was more than that of malignant pathology (96/795 vs. 65/795). In total, 86.8% (269/310) malignant nodules and 93.6% (365/390) benign cases diagnosed by the four TIRADS simultaneously were pathologically confirmed, whereas 8.3% (66/795) of nodules could not be correctly diagnosed by any of the TIRADS, and 12.0% (95/795) nodules could not be consistently diagnosed by all the four TIRADS.

### Discriminant strategy

Most of the 95 nodules were solid, wider-than-tall, without calcifications regardless of pathological diagnosis. Only a small percentage of nodules contained taller-than-wide (1.1%) and microcalcification (8.4%) features. As for echogenic features, benign nodules were predominantly iso/hyperechoic (70.9%), while malignant ones showed predominantly hypoechoic (85.0%). As for margin features, a well-circumscribed margin was predominant in benign nodules (56.4%), and a lobulated or irregular margin was predominant in malignant cases (77.5%) (Supplementary Table 5).

Stepwise discriminant analysis was used to distinguish the variables that best identified benign and malignant nodules. Five commonly used variables were included as predictor variables for malignant thyroid nodules. Stepwise discriminant analysis screened the echogenicity variable ( $F = 34.87$ ,  $p < 0.001$ ). The discriminant function was  $Y = 2.368 \times \text{echogenicity} - 1.421$ . This discriminant function was statistically different (Wilks' lambda = 0.74,  $p < 0.001$ ) and had an excellent predictive value as it could correctly predict the classification of 79.0% of cases.

According to the above discriminant function, the discriminant strategy (DS) based on nodular features was as follows: Iso- or hyper-echogenicity nodules should be considered benign. Hypo- or marked hypo-echogenicity nodules should be regarded as malignant. The diagnostic results of this strategy remained consistent with the above prediction results (Table 2).

**Table 1** Estimated malignant risk of the four TIRADS according to pathological diagnosis

	Total <i>N</i> = 795	Pathological diagnosis		Recommended Malignant risk	Calculated Malignant risk	<i>p</i>
		Benign <i>N</i> = 461	Malignancy <i>N</i> = 334			
<b>ACR-TIRADS</b>						
TR2	127 (15.9%)	127 (27.5%)	0 (0%)	≤ 2%	0%	< 0.001
TR3	201 (25.3%)	190 (41.2%)	11 (3.3%)	< 5%	5.5%	
TR4	154 (19.4%)	103 (22.3%)	51 (15.3%)	5–20%	33.1%	
TR5	313 (39.4%)	41 (9.0%)	272 (81.4%)	> 20%	86.9%	
<b>Kwak-TIRADS</b>						
3	141 (17.7%)	139 (30.1%)	2 (0.6%)	2.0–2.8%	1.4%	< 0.001
4a	226 (28.4%)	215 (46.6%)	11 (3.3%)	3.6–12.7%	4.9%	
4b	80 (10.1%)	54 (11.7%)	26 (7.7%)	6.8–37.8%	32.5%	
4c	276 (34.7%)	48 (10.5%)	228 (68.3%)	21–91.9%	82.6%	
5	72 (9.1%)	5 (1.1%)	67 (20.1%)	88.7–97.9%	93.1%	
<b>C-TIRADS</b>						
CTR2	10 (1.2%)	10 (2.1%)	0 (0%)	0%	0%	< 0.001
CTR3	142 (17.8%)	139 (30.1%)	3 (0.9%)	≤ 2%	2.0%	
CTR4a	249 (31.3%)	224 (48.6%)	25 (7.5%)	2–10%	10.0%	
CTR4b	124 (15.6%)	58 (12.6%)	66 (19.7%)	10–50%	53.2%	
CTR4c	257 (32.4%)	29 (6.3%)	228 (68.2%)	50–90%	88.7%	
CTR5	13 (1.7%)	1 (0.3%)	12 (3.7%)	≥ 90%	92.3%	
<b>EU-TIRADS</b>						
2	7 (0.9%)	7 (1.5%)	0 (0%)	0%	0%	< 0.001
3	323 (40.6%)	315 (68.3%)	8 (2.4%)	2–4%	2.5%	
4	87 (10.9%)	65 (14.1%)	22 (6.6%)	6–17%	25.3%	
5	378 (47.6%)	74 (16.1%)	304 (91.0%)	26–87%	80.4%	

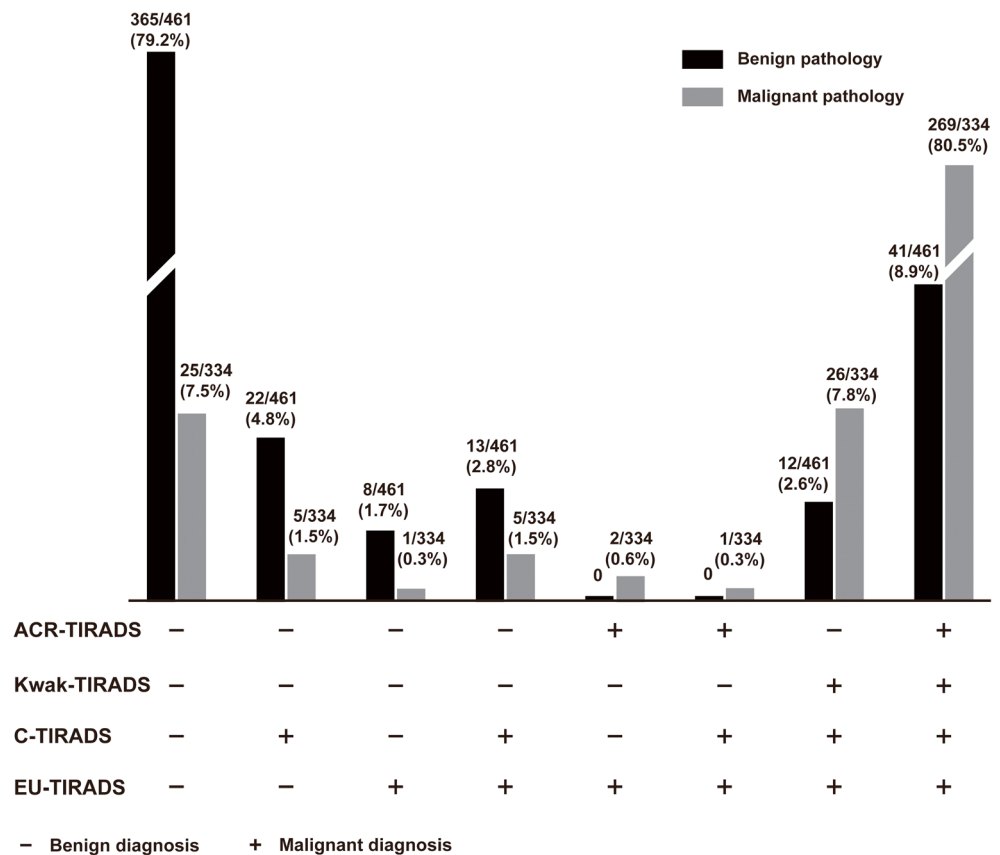
## Performance characteristics

For 95 inconsistently diagnosed nodules screened by at least two TIRADS, DS performed best with an accuracy of 79.0%, followed by Kwak-TIRADS (72.6%) (Fig. 3). Table 3 examines the connection modes of various TIRADS and DS for multiple TIRADS inconsistently diagnosed nodules. Combining DS and ACR-TIRADS in parallel resulted in a significant increase in accuracy (from 61.1 to 80.0%), and the AUC of A-DS was significantly improved (0.817 vs. 0.538) compared to those of ACR-TIRADS alone, while a serial test combining DS and C-TIRADS also resulted in a sharp increase of accuracy (from 47.3 to 76.8%), and the AUC of C-DS method was significantly improved (0.776 vs. 0.535) compared to those of C-TIRADS alone. Regardless of using any combined tests, the AUC of combining DS and EU-TIRADS was substantially higher than that of EU-TIRADS alone (0.700 vs. 0.637, 0.742 vs. 0.637). But the serial test may be preferred because of the higher AUC value and the more balanced sensitivity and specificity values.

Table 4 shows the diagnostic performance of assessment methods built from the screening process, the DS, and combined tests in the overall sample. The sensitivity and AUC were highest for the SP+DS method compared to the four TIRADS (91.3%, 0.895). The specificity was highest for ACR-TIRADS (91.1%), followed by Kwak-TIRADS and the SP+DS method with no significance between them (88.5% vs. 87.6%,  $p > 0.05$ ). When evaluating new methods including combined tests, the sensitivity and AUC were highest for the SP+A+DS method (Parallel mode) (91.6%, 0.896), while the specificity and AUC was highest for the SP+C+DS method (Serial mode) (87.9%, 0.891). For a total of 31 initial Bethesda 3 and 4 nodules (3.9%, 31/795), of which 17 (54.8%) were pathologically malignant and 14 (45.2%) benign, the frequency of correct diagnosis was highest for the SP+DS method and C-TIRADS (both were 20/31), followed by Kwak-TIRADS (19/31) (Supplementary Table 6).

We further examined the performance of one TIRADS combined with another TIRADS or DS in the overall sample (Table 5). Despite a decrease in specificity (from 91.1% to 88.5%), combining ACR-TIRADS and Kwak-TIRADS via

**Fig. 2** Frequency distribution plot of different TIRADS in assessing pathological benign and malignant nodules. The definitions for the malignant and benign results under optimal cut point are as follows: If we set category 5 of ACR-TIRADS as the best cut-off point value, diagnostic malignant nodules are set to be category 5, whereas category 1 to 4 indicates diagnostic benign nodules; If category 5 of EU-TIRADS is set to be best cut-off point value, category 5 indicates diagnostic malignant nodules, while category 1 to 4 represents diagnostic benign nodules; If we set category 4b of C-TIRADS as best cut-off point value, category 4b or 4c or 5 is set to be diagnostic malignant nodules, whereas category 1 to 4a indicates diagnostic benign nodules. If category 4c of Kwak-TIRADS is set to be the best cut-off point value, category 4c or 5 indicates diagnostic malignant nodules, while category 1 to 4b represents diagnostic benign nodules



parallel test resulted in significant improvements in the sensitivity and AUC compared to ACR-TIRADS (89.2% vs. 81.4%, 0.889 vs. 0.863). Although the p-value is at the boundary for statistical significance for the sensitivity (from 91.0 to 92.5%,  $p = 0.053$ ), combining EU-TIRADS and DS via parallel test resulted in significant improvements in AUC (from 0.875 to 0.882,  $p = 0.0245$ ). There are three ways to improve the specificity of C-TIRADS, including combining with Kwak-TIRADS, EU-TIRADS and DS. Combining C-TIRADS and

DS results in the highest AUC (0.887,  $p = 0.0013$ ), followed by Kwak-TIRADS (0.884,  $p = 0.0062$ ), while the lowest AUC was EU-TIRADS (0.879,  $p = 0.0064$ ).

Figure 4 illustrates recommended strategies to improve ultrasound accuracy based on this article’s findings. For suspicious or indeterminate nodules, it was recommended to use two-TIRADS combined tests or one TIRADS combined with DS. But Kwak-TIRADS could be used alone. If someone considered the inconsistent results of two TIRADS, it was recommended to use the DS directly for judgment.

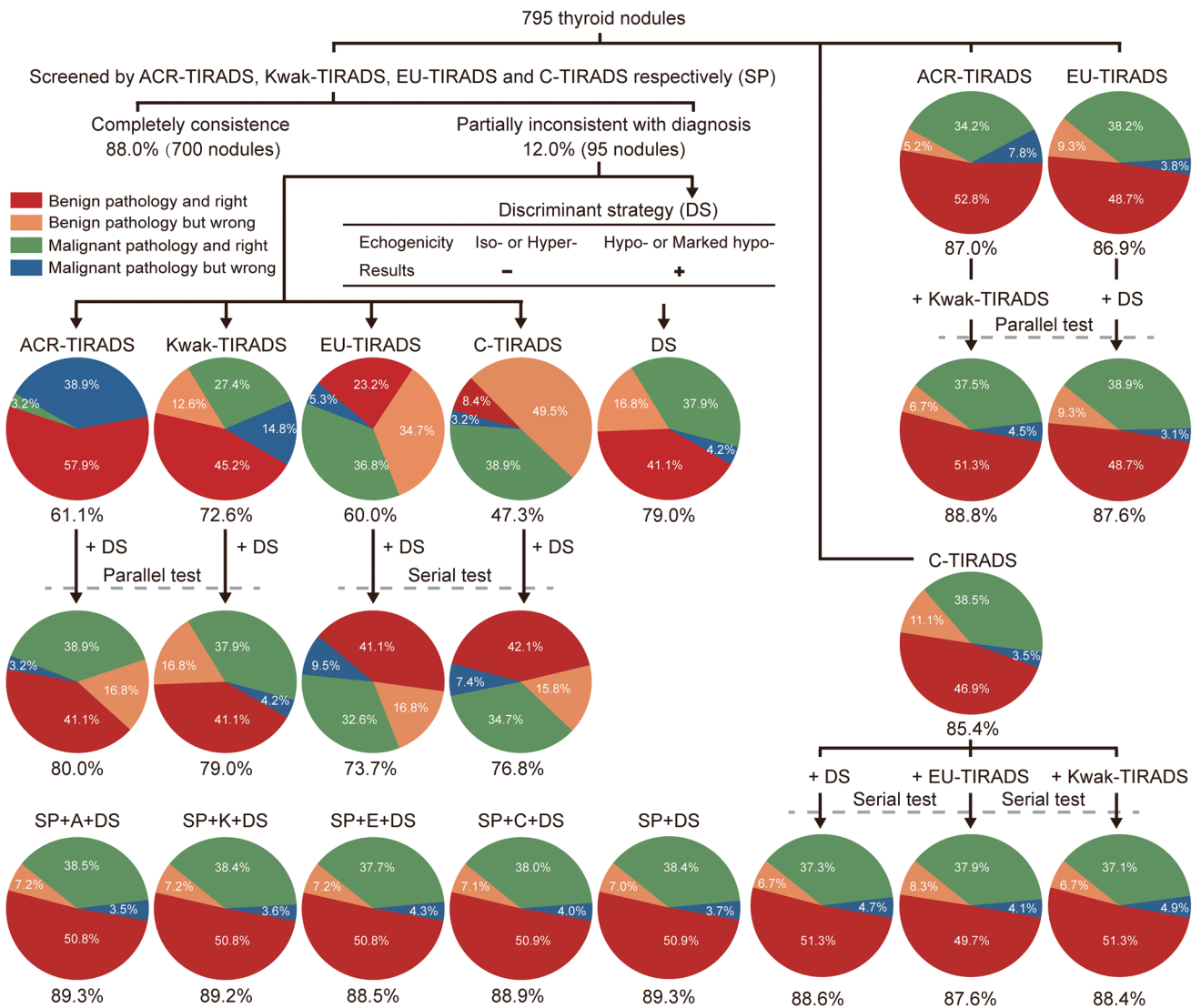
**Table 2** Classification results of discriminant analysis and criteria for differentiating nodules subgroups

Actual	Predicted group membership		
	Benign	Malignance	Total
Benign	39 (70.9%)	16 (29.1%)	55 (100%)
Malignance	4 (10.0%)	36 (90.0%)	40 (100%)
Criteria	Echogenicity		
	Iso- or Hyper-	Hypo-	Marked hypo-
	-	+	+

Note. 79.0% of originally inconsistently grouped cases are correctly classified. Plus means a nodule is diagnosed as malignance and minus represents a benign diagnosis based on the different levels of echogenicity. Marked hypo-echogenicity means the echogenicity is lower than that of the strap muscles of the neck

### Discussion

Our study compared the diagnostic performance of the four TIRADS and showed that all four TIRADS have good diagnostic performance. The four TIRADS screening procedures resulted in 12.0% inconsistently diagnosed nodules. We then tested a strategy focusing on this subgroup of nodules to establish methods for improving diagnostic accuracy. The results showed that established criteria based on the independent variable of echogenicity could fully predict the discriminant results with an accuracy of 79.0%, followed by Kwak-TIRADS (72.6%). The diagnostic performance of the SP+DS method was significantly higher than that of the four



**Fig. 3** Summary of methods and strategies included in the article analysis process. SP: Screening procedures, DS: Discriminant strategy, SP+DS: The evaluation method consists of the four TIRADS screening procedures with partially inconsistently diagnosed nodules judged by discriminant strategy, SP+A/C/K/E+DS: The evaluation method consists of the four TIRADS screening procedures with partially inconsistently diagnosed nodules judged by ACR-TIRADS/C-TIRADS/Kwak-TIRADS/EU-TIRADS and combined with discriminant strategy. The parallel test is defined as follows: The same nodule is defined as

benign only when both tests are diagnosed as benign, or malignance when one of the tests is diagnosed as malignance. The serial test is defined as follows: The same nodule is defined as malignance only when both tests are diagnosed as malignance, or benign when one of tests is diagnosed as benign. The dotted line indicates that the two longitudinal TIRADS or TIRADS and discriminant strategy (DS) are combined in parallel or serial. The numbers at the bottom of the pie chart represent accuracy

TIRADS. Especially, the four TIRADS can substantially improve the diagnostic results of partially inconsistently diagnosed nodules when combined with DS, thus improving the diagnostic performance of the constructed methods (including SP+A+DS, SP+K+DS, SP+E+DS, and SP+C+DS). When the DS was applied to the overall sample, significant improvements in the diagnostic performance of C-TIRADS and EU-TIRADS could be obtained by combining tests. Two-TIRADS parallel or serial tests can also help improve the diagnostic performance of ACR-TIRADS and C-TIRADS by combining Kwak-TIRADS.

In this study, the sensitivity of the four TIRADS ranged from 81.4 to 91.6%, specificity from 80.9 to 91.1%, and AUC 0.863 to 0.884, which indicates that all TIRADS have a good diagnostic performance. C-TIRADS has the highest sensitivity, while ACR-TIRADS has the highest specificity, consistent with the results of previous studies [16–18]. The classification screening results corroborated our hypothesis that the partially inconsistently diagnosed results of the four TIRADS for some nodules are precisely the reason for their differential diagnostic performance.

**Table 3** Diagnostic performance of different TIRADS combined with discriminant strategy using parallel or serial tests on partially inconsistently diagnosed nodules subgroups

<i>N</i> = 95	Sensitivity (%)	Specificity (%)	AUC (95% CI)	<i>p</i>
To increase sensitivity (parallel test)				
A-DS	92.5	70.9	0.817 (0.725–0.889)	< 0.001
A	7.5	100.0	0.538 (0.432–0.640)	
E-DS	100.0	40.0	0.700 (0.597–0.790)	0.0183
E	87.5	40.0	0.637 (0.532–0.734)	
C-DS	100.0	12.7	0.564 (0.458–0.665)	0.2160
C	92.5	14.6	0.535 (0.430–0.638)	
K-DS	90.0	70.9	0.805 (0.710–0.879)	0.0227
K	65.0	78.2	0.716 (0.614–0.804)	
To increase specificity (serial test)				
A-DS	5.0	100.0	0.525 (0.420–0.628)	0.3173
A	7.5	100.0	0.538 (0.432–0.640)	
E-DS	77.5	70.9	0.742 (0.642–0.826)	0.0082
E	87.5	40.0	0.637 (0.532–0.734)	
C-DS	82.5	72.7	0.776 (0.679–0.855)	< 0.001
C	92.5	14.6	0.535 (0.430–0.638)	
K-DS	65.0	78.2	0.716 (0.614–0.804)	1.0000
K	65.0	78.2	0.716 (0.614–0.804)	

A, ACR-TIRADS; E, EU-TIRADS; C, C-TIRADS; K, Kwak-TIRADS; DS, discriminant strategy; AUC, area under the curve; CI, confidence interval. The parallel test is defined as follows: The same nodule is defined as benign only when both TIRADS and DS are diagnosed as benign, or defined as malignance when one of tests is diagnosed as malignance. The serial test is defined as follows: The same nodule is defined as malignance only when both TIRADS and DS are diagnosed as malignance, or defined as benign when one of tests is diagnosed as benign

Without adding other new indicators, we used discriminant analysis to screen out a predictor- echo characteristics. The SP+DS method constructed achieved better diagnostic performance. One could argue whether these malignant indicators depend on the probability distribution of the sample. It must be noted that the remaining nodules partially inconsistent with the diagnosis screened by the four TIRADS are less likely to show highly malignant features such as taller-than-wide, microcalcifications, and infiltrative margins. On the contrary, most of them show less highly malignant features such as solid, hypo-echogenicity, irregular margin and macrocalcifications. Moreover, it can be observed that nodules with highly malignant features often have multiple malignant features simultaneously and are more likely to be correctly diagnosed by various TIRADS [4, 6, 7, 12].

The new evaluation methods have their advantages. The consistent diagnosis of the four TIRADS can provide immediate feedback to increase confidence in confirming the diagnosis. What is more, the false-positive or false-negative rate could be

effectively controlled and balanced, reducing the rate of unnecessary punctures and improving diagnostic accuracy, which is currently two essential goals in nodular diagnosis [19]. In addition, the DS has been simplified and easy to master. Considering the ease of use, it is recommended that in practice the four TIRADS screening procedures are best carried out with the help of structured forms or designed procedures. Most importantly, the diagnostic performance of the SP+DS method and other SP-based methods have been improved compared with the four TIRADS. As for the Bethesda 3 and 4 nodules, the correct diagnostic frequency of the SP+DS method was even the highest. However, due to the limited sample size, the latter conclusion must be confirmed in future studies.

We further explored the clinical applicability of the DS and extended it to the overall sample. The results showed that combined modes between the DS and the four TIRADS differed in the partially inconsistently diagnosed nodules. Those TIRADS with high sensitivity, such as EU-TIRADS and C-TIRADS, are applicable to the serial test to improve specificity, while ACR-TIRADS and Kwak-TIRADS with better specificity performance are suitable for the parallel test. These results suggest that the way the DS is applied depends on the different characteristics of each TIRADS [19, 20]. The situation seems to be somewhat different in the overall sample. For example, EU-TIRADS with high sensitivity seemed to be able to continue to improve sensitivity through the parallel test. It should be noted that various kinds of TIRADS combine the DS in different ways, which may also reflect differences in the weighting of echo characteristics in the various TIRADS. For C-TIRADS, both in partially inconsistently diagnosed nodules and in the sample overall, C-TIRADS and DS are combined using a serial test to improve specificity, which may be attributed to the fact that hypo-echogenicity is not a highly malignant risk feature in its lexicon and many nodules diagnosed as malignance by C-TIRADS with ill-defined margin were correctly diagnosed as benign according to the serial strategy [12, 17].

Notably, in our study, Kwak-TIRADS has a good balance in terms of sensitivity and specificity, which is consistent with previous studies [20, 21]. For partially inconsistently diagnosed nodules, Kwak-TIRADS also exhibited the best performance besides the SP+DS method. In the overall sample, ACR-TIRADS in parallel combined with Kwak-TIRADS reduced the false negative rate, and C-TIRADS in serial combined with Kwak-TIRADS reduced the false positive rate. However, although C-TIRADS can be combined with Kwak-TIRADS to improve specificity, the accuracy is the same as Kwak-TIRADS (both 88.4%), so it is recommended to use Kwak-TIRADS directly. Taken together, it may not be necessary to combine another strategy to achieve better diagnostic performance for Kwak-TIRADS.



**Table 4** Diagnostic performance of evaluation methods consisting of the discriminant strategy alone or combined with the four TIRADS using serial or parallel tests after the screening procedures

<i>N</i> = 795	Sensitivity (%)	Specificity (%)	AUC (95% CI)
A	81.4 (76.8–85.5) <sup>a**</sup>	91.1 (88.1–93.5) <sup>a**</sup>	0.863 (0.837–0.886) <sup>a**</sup>
K	88.3 (84.4–91.6) <sup>a**</sup>	88.5 (85.2–91.3)	0.884 (0.860–0.906) <sup>a*</sup>
C	91.6 (88.1–94.4)	80.9 (77.0–84.4) <sup>a**</sup>	0.863 (0.837–0.886) <sup>a**</sup>
E	91.0 (87.4–93.9)	84.0 (80.3–87.2) <sup>a**</sup>	0.875 (0.850–0.897) <sup>a**</sup>
SP+DS	91.3 (87.8–94.1)	87.6 (84.3–90.5)	0.895 (0.871–0.915)
Parallel test			
SP+A+DS	91.6 (88.1–94.4) <sup>b**</sup>	87.6 (84.3–90.5) <sup>b**</sup>	0.896 (0.873–0.917) <sup>b**</sup>
SP+K+DS	91.3 (87.8–94.1) <sup>c**</sup>	87.6 (84.3–90.5)	0.895 (0.871–0.915) <sup>c*</sup>
Serial test			
SP+C+DS	90.4 (86.7–93.4)	87.9 (84.5–90.7) <sup>d**</sup>	0.891 (0.868–0.912) <sup>d**</sup>
SP+E+DS	89.8 (86.1–92.8)	87.6 (84.3–90.5) <sup>e**</sup>	0.887 (0.863–0.908) <sup>e*</sup>

A, ACR-TIRADS; E, EU-TIRADS; C, C-TIRADS; K, Kwak-TIRADS; AUC, area under the curve; CI, confidence interval; SP, screening procedures; DS, discriminant strategy; SP+DS, the evaluation method consists of the four TIRADS screening procedures with partially inconsistently diagnosed nodules judged by discriminant strategy; SP+A/K/C/E+DS, the evaluation method consists of the four TIRADS screening procedures with partially inconsistently diagnosed nodules judged by ACR-TIRADS/Kwak-TIRADS/C-TIRADS/EU-TIRADS and combined with discriminant strategy. The parallel test is defined as follows: The same nodule is defined as benign only when both TIRADS and DS are diagnosed as benign, or malignance when one of tests is diagnosed as malignance. The serial test is defined as follows: The same nodule is defined as malignance only when both TIRADS and DS are diagnosed as malignance, or benign when one of tests is diagnosed as benign. <sup>a</sup> SP+DS is the statistical control group, <sup>b</sup> ACR-TIRADS is the statistical control group, <sup>c</sup> Kwak-TIRADS is the statistical control group, <sup>d</sup> C-TIRADS is the statistical control group, <sup>e</sup> EU-TIRADS is the statistical control group. \*  $p < 0.05$ , \*\*  $p < 0.01$

In the clinical setting, two or more evaluation systems are usually considered for suspicious or indeterminate nodules with few highly malignant features [22]. Considering that the customary use of TIRADS may differ among institutions or individuals, diagnostic combinations that are both accurate and clinically significant need to be examined. Based on the results of this study, if a nodule is suspicious or uncertain diagnosis, two TIRADS or the combination of one TIRADS with DS using a parallel or serial tests can be considered to help improve the accuracy, where Kwak-TIRADS can be directly used without the combination test. On the other hand, the diagnostic consistency of the two TIRADS at the optimal cut point can be examined. If the results of the two selected TIRADS are inconsistent, considering the time cost of the screening process, it is suggested to directly use the DS, which can significantly improve the accuracy.

ACR-TIRADS is a commonly used TIRADS with high specificity, effectively reducing unnecessary FNA rates [17, 18, 23]. But false negatives are a concern. According to the results of this study, we may suggest using a parallel test combined with Kwak-TIRADS for judgment to obtain a balance of sensitivity and specificity. However, as with ACR-TIRADS, there seems to be value in the uneven diagnostic performance of TIRADS [19]. Despite using the serial strategy, some new methods' specificity does not seem to exceed

that of ACR-TIRADS. On the contrary, with the combining strategy, this study has obtained multiple sets of assessment methods with a balance of sensitivity and specificity, even some methods that can enhance sensitivity, e.g., A-DS and K-DS. Whether these methods have clinical application need to be further investigated.

Our study has some limitations. First, all patients in this study with malignant thyroid tumors were confined to PTC. Whether the conclusion of this study applies to other thyroid malignant tumors needs to be verified. Second, the selection of optimal cut-off points for TIRADS, especially considering the balance of sensitivity and specificity, might affect the results of this study. However, the cut-off point of each TIRADS is relatively stable. The ACR-TIRADS and EU-TIRADS are mostly set at category 4 or 5, while the Kwak-TIRADS is set chiefly at category 4c to maximize the balance of sensitivity and specificity to ensure the accuracy of diagnosis [19, 24–26]. As mentioned above, the screening procedures almost exclude nodules with multiple highly malignant features, so it can be predicted that, even though samples might be different, similar criteria may still be obtained according to this research strategy. Further research with larger samples or the other thyroid carcinoma is needed to confirm the above hypothesis. In addition, it must be acknowledged that classification below the optimal cut

**Table 5** Diagnostic performance of the discriminant strategy combined with TIRADS or two combined TIRADS using parallel or serial combination strategies

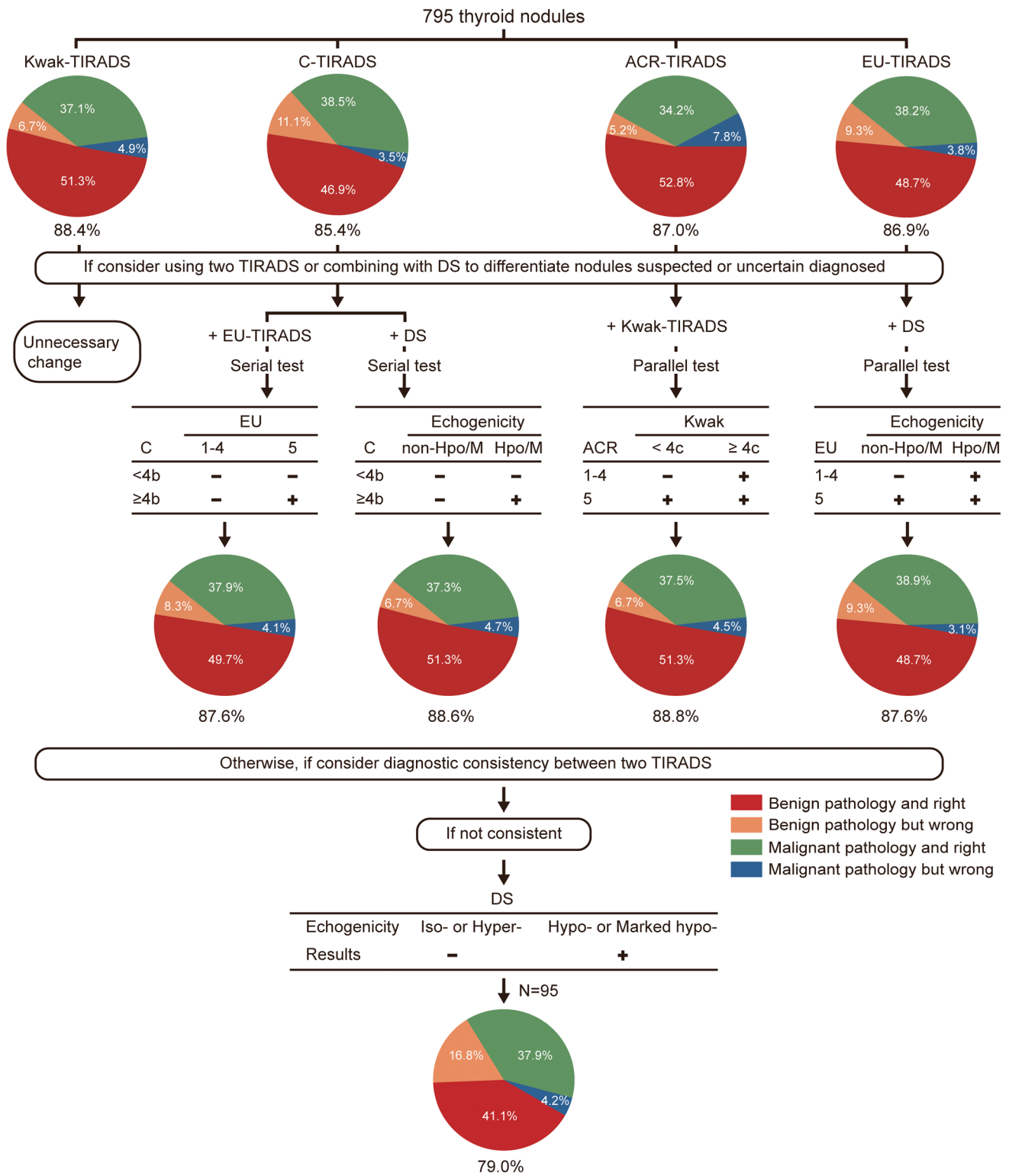
<i>N</i> = 795	Sensitivity (%)	Specificity (%)	AUC (95% CI)	Note
A	81.4 (76.8–85.5)	91.1 (88.1–93.5)	0.863 (0.837–0.886)	-
K	88.3 (84.4–91.6)	88.5 (85.2–91.3)	0.884 (0.860–0.906)	-
C	91.6 (88.1–94.4)	80.9 (77.0–84.4)	0.863 (0.837–0.886)	-
E	91.0 (87.4–93.9)	84.0 (80.3–87.2)	0.875 (0.850–0.897)	-
To increase sensitivity (parallel test)				
A-K	89.2 (85.4–92.3)	88.5 (85.2–91.3)	0.889 (0.865–0.910)	Sig
<i>p</i>	< 0.001 <sup>a</sup>	< 0.001 <sup>a</sup>	0.0016 <sup>a</sup>	
A-C	92.2 (88.8–94.9)	80.9 (77.0–84.4)	0.866 (0.840–0.889)	NS
A-E	91.0 (87.4–93.9)	84.0 (80.3–87.2)	0.875 (0.850–0.897)	NS
K-C	91.6 (88.1–94.4)	80.9 (77.0–84.4)	0.863 (0.837–0.886)	NS
K-E	91.0 (87.4–93.9)	84.0 (80.3–87.2)	0.875 (0.850–0.897)	NS
E-C	92.5 (89.1–95.1)	79.2 (75.2–82.8)	0.858 (0.832–0.882)	NS
A-DS	95.8 (93.1–97.7)	79.8 (75.9–83.4)	0.878 (0.853–0.900)	NS
K-DS	95.5 (92.7–97.5)	79.6 (75.6–83.2)	0.876 (0.851–0.898)	NS
C-DS	92.5 (89.1–95.1)	80.7 (76.8–84.2)	0.866 (0.840–0.889)	NS
E-DS	92.5 (89.1–95.1)	84.0 (80.3–87.2)	0.882 (0.858–0.897)	Sig
<i>p</i>	0.053 <sup>b</sup>	1.000 <sup>b</sup>	0.0245 <sup>b</sup>	
To increase specificity (serial test)				
A-K	80.5 (75.9–84.6)	91.1 (88.1–93.5)	0.858 (0.832–0.882)	NS
A-C	80.8 (76.2–84.9)	91.1 (88.1–93.5)	0.860 (0.834–0.883)	NS
A-E	81.4 (76.8–85.5)	91.1 (88.1–93.5)	0.863 (0.837–0.886)	NS
K-C	88.3 (84.4–91.6)	88.5 (85.2–91.3)	0.884 (0.860–0.906)	Sig
<i>p</i>	0.001 <sup>c</sup>	< 0.001 <sup>c</sup>	0.0062 <sup>c</sup>	
K-E	88.3 (84.4–91.6)	88.5 (85.2–91.3)	0.884 (0.860–0.906)	NS
E-C	90.1 (86.4–93.1)	85.7 (82.1–88.8)	0.879 (0.854–0.901)	Sig
<i>p</i>	0.064 <sup>d</sup>	< 0.001 <sup>d</sup>	0.0064 <sup>d</sup>	
A-DS	81.1 (76.5–85.2)	91.1 (88.1–93.5)	0.861 (0.835–0.885)	NS
K-DS	88.3 (84.4–91.6)	88.5 (85.2–91.3)	0.884 (0.860–0.906)	NS
C-DS	88.9 (85.1–92.1)	88.5 (85.2–91.3)	0.887 (0.863–0.908)	Sig
<i>p</i>	0.004 <sup>e</sup>	< 0.001 <sup>e</sup>	0.0013 <sup>e</sup>	
E-DS	88.3 (84.4–91.6)	88.3 (85.0–91.1)	0.883 (0.859–0.905)	NS

A, ACR-TIRADS; E, EU-TIRADS; C, C-TIRADS; K, Kwak-TIRADS; AUC, area under the curve; CI, confidence interval; NS, no significance; Sig, Significance. The parallel test is defined as follows: The same nodule is defined as benign only when both tests are diagnosed as benign, or malignance when one of tests is diagnosed as malignance. The serial test is defined as follows: The same nodule is defined as malignance only when both tests are diagnosed as malignance, or benign when one of tests is diagnosed as benign. <sup>a</sup> ACR-TIRADS is the statistical control group, <sup>b</sup> EU-TIRADS is the statistical control group, <sup>c, d, e</sup> C-TIRADS is the statistical control group

point also has the risk of malignancy, and it is still necessary to consider whether to carry out an FNA examination based on the size of nodules, personal or family history of cancer, and changes in nodules during the follow-up period. Third, as mentioned in the “FNA, cytopathology and histopathology” section, repeated FNAs are not routinely performed in our institution, which may cause selection bias. What is more, there is still a 1–2% false-negative rate based on repeated results, which might overestimate benign nodules and affect the diagnostic performance of each TIRADS.

Further study could use benign surgical pathology results to exclude this potential bias [1, 15, 27].

In conclusion, it is undeniable that various TIRADS have good diagnostic performance, but how to further improve the diagnostic accuracy is a question worth exploring. This study is the first to analyze and compare in detail the misdiagnosed and missed cases of different TIRADS. We explored new methods without additional diagnostic indicators and achieved an effective improvement in accuracy. The recommended strategies our findings provide may help to improve



**Fig. 4** Summary of recommended strategies to improve ultrasound accuracy based on this article’s findings. ACR: ACR-TIRADS, EU: EU-TIRADS, C: C-TIRADS, Kwak: Kwak-TIRADS, DS: Discriminant strategy, Hpo/M: Nodules with hypo- or marked hypo-

echogenicity, non-Hpo/M: Nodules with iso- or hyper-echogenicity. The best diagnostic cut-off values of ACR-TRADS, Kwak-TIRADS, C-TIRADS, and EU-TIRADS are TR5, 4c, CTR 4b, and grade 5 in this article. The numbers at the bottom of the pie chart represent accuracy

the diagnostic accuracy of ultrasound uncertain or suspicious nodules.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00330-022-09126-2>.

**Funding** This work was supported by the Science and Technology Plan of Guangzhou under Grant No. 202103000048.

## Declarations

**Guarantor** The scientific guarantor of this publication is Jian Kuang.

**Conflict of interest** The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

**Statistics and biometry** One of the authors (Jian Kuang) has significant statistical expertise.

**Informed consent** Written informed consent was waived by the Institutional Review Board.

**Ethical approval** Institutional Review Board approval was obtained (KY-Q-2022-138-02).

## Methodology

- retrospective
- diagnostic or prognostic study / observational
- performed at one institution

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Gharib H, Papini E, Garber JR et al (2016) 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* 22:622–639
- Brito JP, Gionfriddo MR, Al Nofal A et al (2014) The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. *J Clin Endocrinol Metab* 99:1253–1263
- Horvath E, Majlis S, Rossi R et al (2009) An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. *J Clin Endocrinol Metab* 94:1748–1751
- Kwak JY, Han KH, Yoon JH et al (2011) Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. *Radiology* 260:892–899
- Haugen BR, Alexander EK, Bible KC et al (2016) 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* 26:1–133
- Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L (2017) European Thyroid Association Guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *Eur Thyroid J* 6:225–237
- Tessler FN, Middleton WD, Grant EG et al (2017) ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee. *J Am Coll Radiol* 14:587–595
- Shin JH, Baek JH, Chung J et al (2016) Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. *Korean J Radiol* 17:370–395
- Leni D, Seminati D, Fior D et al (2021) Diagnostic performances of the ACR-TIRADS system in thyroid nodules triage: a prospective single center study. *Cancers* 13:2230
- Rossi ED, Pantanowitz L, Raffaelli M, Fadda G (2021) Overview of the ultrasound classification systems in the field of thyroid cytology. *Cancers* 13:3133
- Ruan JL, Yang HY, Liu RB et al (2019) Fine needle aspiration biopsy indications for thyroid nodules: compare a point-based risk stratification system with a pattern-based risk stratification system. *Eur Radiol* 29:4871–4878
- Zhou J, Yin L, Wei X et al (2020) 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. *Endocrine* 70:256–279
- Delellis RA, Lloyd RV, Heitz PU, Eng C (2004) Pathology and genetics of tumors of endocrine organs, 1st edn. IARC Press, International Agency for Research on Cancer, Lyon
- Cibas ES, Ali SZ (2009) The Bethesda System for Reporting Thyroid Cytopathology. *Am J Clin Pathol* 132:658–665
- Kim JH, Baek JH, Lim HK et al (2018) 2017 Thyroid radiofrequency ablation guideline: Korean Society of Thyroid Radiology. *Korean J Radiol* 19:632–655
- Chen Q, Lin M, Wu S (2022) Validating and comparing C-TIRADS, K-TIRADS and ACR-TIRADS in stratifying the malignancy risk of thyroid nodules. *Front Endocrinol (Lausanne)* 13:899575
- Zhou J, Song Y, Zhan W et al (2021) Thyroid imaging reporting and data system (TIRADS) for ultrasound features of nodules: multicentric retrospective study in China. *Endocrine* 72:157–170
- Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH (2021) Unnecessary thyroid nodule biopsy rates under four ultrasound risk stratification systems: a systematic review and meta-analysis. *Eur Radiol* 31:2877–2885
- Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH (2020) Diagnostic performance of four ultrasound risk stratification systems: a systematic review and meta-analysis. *Thyroid* 30:1159–1168
- Yoon JH, Lee HS, Kim EK, Moon HJ, Park VY, Kwak JY (2020) Pattern-based vs. score-based guidelines using ultrasound features have different strengths in risk stratification of thyroid nodules. *Eur Radiol* 30:3793–3802
- Koh J, Kim SY, Lee HS et al (2018) (2018) diagnostic performances and interobserver agreement according to observer

- experience: a comparison study using three guidelines for management of thyroid nodules. *Acta Radiol* 59:917–923
22. Hoang JK, Asadollahi S, Durante C, Hegedüs L, Papini E, Tessler FN (2022) An international survey on utilization of five thyroid nodule risk stratification systems: a needs assessment with future implications. *Thyroid* 32:675–681
  23. Huh S, Yoon JH, Lee HS, Moon HJ, Park VY, Kwak JY (2021) Comparison of diagnostic performance of the Acr and Kwak Tirads applying the Acr Tirads' size thresholds for Fna. *Eur Radiol* 31: 5243–5250
  24. Gao L, Xi X, Jiang Y et al (2019) Comparison among TIRADS (ACR TI-RADS and KWAK- TI-RADS) and 2015 ATA Guidelines in the diagnostic efficiency of thyroid nodules. *Endocrine* 64:90–96
  25. Schenke S, Klett R, Seifert P, Kreissl MC, Görge R, Zimny M (2020) Diagnostic performance of different thyroid imaging reporting and data systems (Kwak-TIRADS, EU-TIRADS and ACR TI-RADS) for risk stratification of small thyroid nodules ( $\leq 10$  mm). *J Clin Med* 9:236
  26. Xu T, Wu Y, Wu RX et al (2019) Validation and comparison of three newly-released Thyroid Imaging Reporting and Data Systems for cancer risk determination. *Endocrine* 64:299–307
  27. Hahn SY, Shin JH, Han BK, Ko EY, Ko ES (2013) Ultrasonography-guided core needle biopsy for the thyroid nodule: does the procedure hold any benefit for the diagnosis when fine-needle aspiration cytology analysis shows inconclusive results? *Br J Radiol* 86:20130007

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.