World Journal of Radiology

Submit a Manuscript: https://www.f6publishing.com

World J Radiol 2022 January 28; 14(1): 19-29

DOI: 10.4329/wjr.v14.i1.19

ISSN 1949-8470 (online)

ORIGINAL ARTICLE

Retrospective Cohort Study

Effect of training on resident inter-reader agreement with American College of Radiology Thyroid Imaging Reporting and Data System

Yang Du, Meredith Bara, Prayash Katlariwala, Roger Croutze, Katrin Resch, Jonathan Porter, Medica Sam, Mitchell P Wilson, Gavin Low

ORCID number: Yang Du 0000-0002-9150-2932; Meredith Bara 0000-0002-0565-050X; Prayash Katlariwala 0000-0002-5822-1071; Roger Croutze 0000-0002-0060-2170; Katrin Resch 0000-0001-6654-3153; Jonathan Porter 0000-0002-8067-1722; Medica Sam 0000-0002-3962-633X; Mitchell P Wilson 0000-0002-1630-5138; Gavin Low 0000-0002-4959-8934.

Author contributions: Du Y, Bara M and Low G designed the study; Du Y, Bara M, Croutze R, Resch K, Porter J, Sam M, Wilson MP and Low G performed the research; Du Y, Bara M, Katlariwala P, Low G and Wilson MP analyzed the data and wrote the manuscript; all authors have read and approved the final manuscript.

Institutional review board statement: This retrospective, single-institution observational study was approved by the institutional Health Research Ethics Board (Pro 00104708).

Informed consent statement: This study was exempted from obtaining informed consent.

Conflict-of-interest statement: The authors have no conflict of interest to declare.

Data sharing statement: The raw

Yang Du, Meredith Bara, Prayash Katlariwala, Roger Croutze, Katrin Resch, Jonathan Porter, Medica Sam, Mitchell P Wilson, Gavin Low, Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton T6G 2B7, Alberta, Canada

Corresponding author: Yang Du, BSc, FRCPC, MD, Doctor, Staff Physician, Department of Radiology and Diagnostic Imaging, University of Alberta, 2A2.41 WMC, 8440-112 St NW, Edmonton T6G 2B7, Alberta, Canada. yang.du@usask.ca

Abstract

BACKGROUND

The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) was introduced to standardize the ultrasound characterization of thyroid nodules. Studies have shown that ACR-TIRADS reduces unnecessary biopsies and improves consistency of imaging recommendations. Despite its widespread adoption, there are few studies to date assessing the inter-reader agreement amongst radiology trainees with limited ultrasound experience. We hypothesize that in PGY-4 radiology residents with no prior exposure to ACR TI-RADS, a statistically significant improvement in inter-reader reliability can be achieved with a one hour training session.

AIM

To evaluate the inter-reader agreement of radiology residents in using ACR TI-RADS before and after training.

METHODS

A single center retrospective cohort study evaluating 50 thyroid nodules in 40 patients of varying TI-RADS levels was performed. Reference standard TI-RADS scores were established through a consensus panel of three fellowship-trained staff radiologists with between 1 and 14 years of clinical experience each. Three PGY-4 radiology residents (trainees) were selected as blinded readers for this study. Each trainee had between 4 to 5 mo of designated ultrasound training. No trainee had received specialized TI-RADS training prior to this study. Each of the readers independently reviewed the 50 testing cases and assigned a TI-RADS score to each case before and after TI-RADS training performed 6 wk apart. Fleiss kappa was used to measure the pooled inter-reader agreement. The relative diagnostic performance of readers, pre- and post-training, when compared

dataset is available from the corresponding author at yang.du@usask.ca. Consent for data sharing was not obtained but the presented data are anonymized and risk of identification is low.

STROBE statement: Guidelines of the STROBE statement have been adopted.

Country/Territory of origin: Canada

Specialty type: Radiology, nuclear medicine and medical imaging

Provenance and peer review: Unsolicited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification
Grade A (Excellent): A, A
Grade B (Very good): B
Grade C (Good): 0
Grade D (Fair): 0
Grade E (Poor): 0

Received: October 12, 2021
Peer-review started: October 12, 2021
First decision: December 9, 2021
Revised: December 21, 2021
Accepted: January 11, 2022
Article in press: January 11, 2022
Published online: January 28, 2022

P-Reviewer: Beck-Razi N, Lee KS, Naganuma H
S-Editor: Wang LL
L-Editor: A
P-Editor: Wang LL

against the reference standard.

RESULTS
There were 33 females and 7 males with a mean age of 56.6 ± 13.6 years. The mean nodule size was 19 ± 14 mm (range from 5 to 63 mm). A statistically significant superior inter-reader agreement was found on the post-training assessment compared to the pre-training assessment for the following variables: 1. "Shape" (k of 0.09 [slight] pre-training *vs* 0.67 [substantial] post-training, $P < 0.001$), 2. "Echogenic foci" (k of 0.28 [fair] pre-training *vs* 0.45 [moderate] post-training, $P = 0.004$), 3. 'TI-RADS level' (k of 0.14 [slight] pre-training *vs* 0.36 [fair] post-training, $P < 0.001$) and 4. 'Recommendations' (k of 0.36 [fair] pre-training *vs* 0.50 [moderate] post-training, $P = 0.02$). No significant differences between the pre- and post-training assessments were found for the variables 'composition', 'echogenicity' and 'margins'. There was a general trend towards improved pooled sensitivity with TI-RADS levels 1 to 4 for the post-training assessment while the pooled specificity was relatively high (76.6%-96.8%) for all TI-RADS level.

CONCLUSION
Statistically significant improvement in inter-reader agreement in the assigning TI-RADS level and recommendations after training is observed. Our study supports the use of dedicated ACR TI-RADS training in radiology residents.

Key Words: Thyroid; Thyroid nodule; American College of Radiology Thyroid Imaging Reporting and Data System; Inter-reader agreement; Ultrasound

Core Tip: There is a statistically significant improvement in inter-reader agreement among radiology trainees with limited ultrasound experience using the American College of Radiology Thyroid Imaging Reporting and Data System (TI-RADS) after training for TI-RADS grading and recommendations. This study demonstrates the learnability of TI-RADS in radiology trainees.

Citation: Du Y, Bara M, Katlariwala P, Croutze R, Resch K, Porter J, Sam M, Wilson MP, Low G. Effect of training on resident inter-reader agreement with American College of Radiology Thyroid Imaging Reporting and Data System. *World J Radiol* 2022; 14(1): 19-29
URL: https://www.wjgnet.com/1949-8470/full/v14/i1/19.htm
DOI: https://dx.doi.org/10.4329/wjr.v14.i1.19

INTRODUCTION

Thyroid nodules are detected in more than 50% of healthy individuals with approximately 95% representing asymptomatic incidental nodules[1-3]. Moreover, an increasing number of thyroid nodules are being detected in recent years on account of improved quality and increased frequency of medical imaging[4]. Although most thyroid nodules are benign and do not require treatment, adequate characterization is necessary in order to identify potentially malignant nodules[1-3]. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) was therefore introduced to standardize the ultrasound characterization of thyroid nodules based on 5 morphologic categories (composition, echogenicity, shape, margins, and echogenic foci). A TI-RADS score is obtained to represent the level of suspicion for cancer and further direct the need for follow-up and/or tissue sampling [5]. First published in 2017, ACR TI-RADS has been widely adopted by many centers worldwide. Studies have shown that ACR-TIRADS reduces unnecessary biopsies and improves consistency of imaging recommendations[6,7].

Despite its widespread adoption, there are few studies available to date assessing the inter-reader reliability of TI-RADS amongst radiology trainees with limited ultrasound experience. A single-institutional study performed in China by Teng *et al* [8] evaluated three trainees with less than three months of ultrasound experience,

demonstrating fair to almost perfect agreement amongst readers for TI-RADS categorization, with improved agreement and diagnostic accuracy after training. To our knowledge, no similar inter-reader agreement studies have been performed in North American trainees. The purpose of this study is to evaluate the inter-reader reliability amongst radiology trainees before and after designated TI-RADS training in a North American institution.

MATERIALS AND METHODS

This retrospective, single-institution observational study was approved by the institutional Health Research Ethics Board (Pro 00104708). This study was exempted from obtaining informed consent. A retrospective review of the local Picture Archiving and Communication System (PACS) was performed to identify thyroid ultrasound studies containing thyroid nodules between July 1, 2019 to July 31, 2020. Included cases required at least 1 thyroid nodule (minimal dimension of 5 mm) with both transverse and sagittal still images and cine video recording in at least 1 plane. Nodules with non-diagnostic image quality, incomplete nodule visualization, and absence of a cine clip covering the entirety of the nodule were excluded. The type of ultrasound make, model, or platform were not considered in the selection process.

Eighty consecutive thyroid nodules meeting eligibility criteria were selected by 2 authors (YD, 6 years clinical experience; MB, 3 years clinical experience) from the eligible ultrasound examinations. A single case could include more than one nodule if sufficient imaging was available to meet inclusion criteria for multiple nodules. Still images of each nodule in both transverse and sagittal planes as well as at least 1 cine video clip of the nodule were saved in a teaching file hosted on our institutional Picture Archiving and Communication System. Each nodule and its representative images/cine clips were saved separately. If a single patient had two nodules, the relevant images and cine clips for each nodule were saved as separate case numbers. Of these, 50 cases were allocated into the "testing" group and 30 cases into the "training" group. Non-random group selection was performed to allow an approximately even distribution of TI-RADS categories within each group and to prevent under-representation of any category. A steering committee consisting of 2 authors including the principal investigator (YD, MB) attempted to evenly divide cases of differentiating difficulty equally between "testing" and "training" groups. This variable approach was selected over a pathological gold standard in an attempt to reduce referral bias in the "testing" group, a situation likely encountered by Teng *et al* [8] where 61% (245/400) of included nodules were pathologically malignant. The trainees were blinded to the distribution approach of the "testing" group.

All patient identifiers were removed apart from age and gender. All cases were evaluated by a consensus review of 3 independent fellowship-trained board-certified staff radiologists with between 1 and 14 years of clinical experience each (GL, MW, MS). Any disagreement on the scoring of nodules for the ACR TI-RADS level was resolved by re-review and consensus discussion. Findings on the consensus review were recorded and set as the standard of reference. This approach has been used in other recent inter-reader reliability studies assessing ACR Reporting and Data Systems [9].

Three PGY-4 radiology residents (trainees) were selected as blinded readers for this study. Each trainee had between 4 to 5 mo of designated ultrasound training, in addition to non-designated ultrasound training on other rotations throughout their training. No trainee had received specialized TI-RADS training prior to this study. Each of the readers independently reviewed the 50 testing cases and assigned TI-RADS score to each case. The readers were provided with a summary chart detailing the ACR TI-RADS classification as described in the ACR TI-RADS White Paper and had access to an online TI-RADS calculator (https://tiradscalculator.com) at the time of independent review[5]. The readers were instructed to assign TI-RADS points for each category including composition, echogenicity, shape, margins, echogenic foci, and to determine the TI-RADS level and ACR TI-RADS recommendations. The pre-training responses were entered into an online survey generated *via* Google Forms. Four weeks after the readers had completed the pre-training assessment; a one hour-long teaching session including a Microsoft PowerPoint presentation illustrating important features of ACR TI-RADS was provided to the readers along with a Microsoft Word document summarizing common areas of disagreement in nodule characterization[5]. The teaching session provided a step-by-step review of the 5 main sonographic features used for nodule scoring in ACR TI-RADS: (1) Composition; (2)

Figure 1 A 51-year-old female with a 1.1 cm × 0.9 cm × 0.9 cm right mid pole thyroid nodule. This nodule was classified correctly with perfect concordance by all 3 readers as solid (+ 2 points), hypoechoic (+ 2 points), taller-than-wide (+ 3 points), smooth margins (+ 0 points), and with punctate echogenic foci (+ 3 points). This had a total points of 10 and a Thyroid Imaging Reporting and Data System level of TR5.

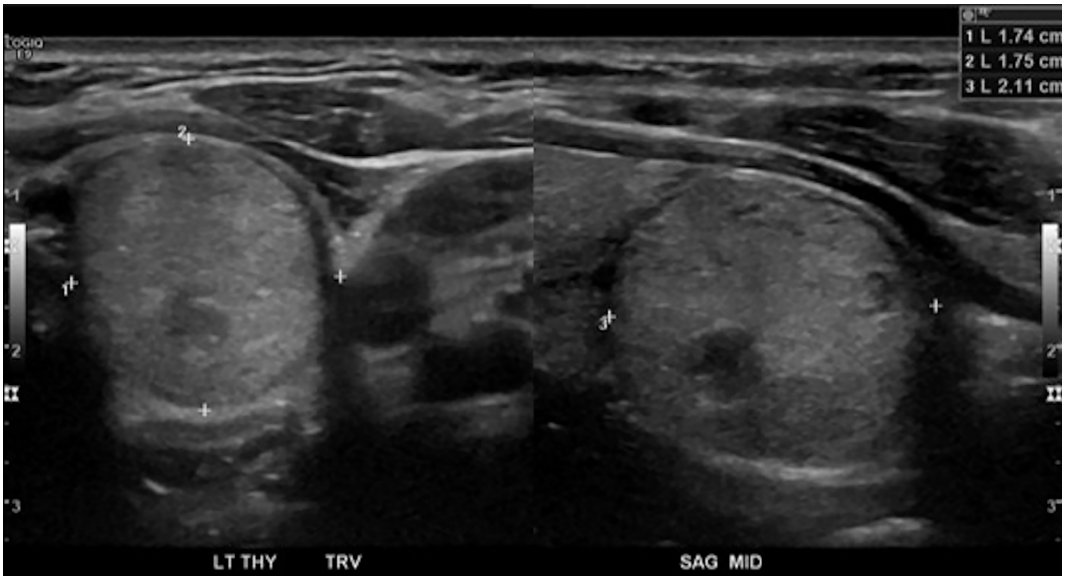

Figure 2 A 45-year-old female with a 1.7 cm × 1.8 cm × 2.1 cm left mid pole thyroid nodule. This nodule was classified by first two readers as Thyroid Imaging Reporting and Data System (TI-RADS) level TR4 and by the third reader as TI-RADS level TR5. The first two readers classified the nodule as solid (+ 2 points), isoechoic (+ 2 points), taller-than-wide (+ 3 points), smooth margins (+ 0 points) and with no echogenic foci (+ 0 points) for a total points of 6 and a TI-RAD level of TR4. For the third reader, a single discrepancy in the scoring of echogenicity as hypoechoic (+ 2 points) rather than isoechoic (+ 1 point) as in the other 2 readers, resulted in a total points of 7 and a TI-RADS level of TR5. As can be seen in the images, the nodule has mixed echogenicity although most of the nodule is isoechoic making this the preferred option.

Echogenicity; (3) Shape; (4) Margin; and (5) Echogenic foci. Each feature's description and interpretation was discussed and illustrated by examples. The readers were given ample opportunity to ask questions, and the consensus panel provided focused clarification to readers in areas of reader uncertainty. Additionally, the trainees were instructed to review the training file that contained the 30 training cases on PACS and corresponding answers were provided for each case. Two weeks after the training session (six weeks after the pre-training assessment), the 50 anonymized cases from the ''testing'' group were re-sent to the readers for independent review. Readers were instructed to re-score the 50 cases and the post-training responses were entered into an online survey generated *via* Google Forms.

Statistical analysis

Categorical variables were expressed as values and percentages. Continuous variables were expressed as the mean ± SD. The following statistical tests were used:

Fleiss kappa (overall agreement) was used to calculate the pooled inter-reader agreement. The kappa (K) value interpretation as suggested by Cohen was used: ≤ 0.20 (slight agreement), 0.21–0.40 (fair agreement), 0.41–0.60 (moderate agreement), 0.61–0.80 (substantial agreement), and 0.81–1.00 (almost perfect agreement)[10].

Paired *t*-test was used to evaluate for significant difference between agreement coefficients[11].

Using the consensus panel as the reference standard, the relative diagnostic parameters (sensitivity, specificity, positive predictive value and negative predictive value) per TI-RADS level were calculated for individual readers and on a pooled basis.

RESULTS

The testing cases comprised of 50 nodules in 40 patients. There were 33 (82.5%) females and 7 males. The mean patient age was 56.6 ± 13.6 years with an age range from 29 to 80 years. Of the 50 nodules, 31 (62%) were located in the right lobe, 18 (36%) in the left lobe and 1 (2%) in the isthmus. The mean nodule size was 19 ± 14 mm with a range from 5 to 63 mm. According to the reference standard that consisted of a consensus panel of 3 fellowship trained staff radiologists, there were 11 (22%) TI-RADS level 1 nodules, 9 (18%) TI-RADS level 2 nodules, 9 (18%) TI-RADS level 3 nodules, 13 (26%) TI-RADS level 3 nodules, and 8 (16%) TI-RADS level 5 nodules.

The pooled inter-reader agreement with the reference standard, pre- and post-training, is listed in Table 1. A statistically significant improvement in reader agreement was demonstrated in post-training inter-reader agreement for nodule shape ($P < 0.001$), presence of echogenic foci ($P = 0.004$), TI-RADS level ($P < 0.001$) and overall recommendation ($P = 0.02$). Each of these categories improved at least one category of agreement. Only margin characterization remained at slight agreement after training. Similarly, the percentage reader agreement with the reference standard for sonographic features (Table 2), TI-RADS levels (Table 3) and recommendations (Table 4) are also included. Figure 1 provides an illustrated example of complete reader concordance for nodule scoring using ACR TI-RADS. In contrast, Figure 2 provides an illustrated example where there is discordance in reader scoring using ACR TI-RADS.

Finally, the relative diagnostic performance of readers, pre- and post-training, when compared against the reference standard is included in Table 5 and Table 6, respectively. Pre-training pooled sensitivities ranged from 22.3%-66.7% and pooled specificity ranged from 72.2%-95.1%, dependent on TI-RADS category. Post-training pooled sensitivities ranged from 40.7%-63% and pooled specificity ranged from 76.6%-96.8%, dependent on TI-RADS category.

DISCUSSION

The overall inter-reader agreement for ACR TI-RADS should take into account the inter-reader agreement of its two major outcome variables – '*TI-RADS level*' and '*ACR TI-RADS recommendations*'. In our study, the inter-reader agreement for *'TI-RADS level'* showed a significant improvement with training ($k = 0.14$ (slight) on the pre-training assessment *vs* $k = 0.36$ (fair) on the post-training assessment)[12]. Our inter-reader agreement for *'ACR TI-RADS recommendations'* also showed a significant improvement with training ($k = 0.36$ (fair) on the pre-training assessment *vs* $k = 0.50$ (moderate) on the post-training assessment [P = 0.02]). Our findings suggest that even a single didactic training session can significantly improve the overall inter-reader agreement in radiology residents. Our findings compare favorably with other inter-reader agreement studies involving ACR TI-RADS. A study by Hoang *et al*[7] involving 8 board certified radiologists (2 from academic centers with subspecialty training in US and 6 from private practice with no subspecialty training in US) found a fair ($k = 0.35$) inter-reader agreement for *'TI-RADS level'*, and moderate ($k = 051$) inter-reader agreement for *'ACR TI-RADS recommendations'*[7]. Teng *et al*[8] assessed the learnability and reproducibility of ACR TI-RADS in post-graduate freshmen. The study included 3 readers with < 3 mo ultrasound experience and 3 experts with > 15 years ultrasound experience each. The readers independently evaluated 4 groups of nodules

Table 1 Pooled inter-reader agreement with the reference standard

	Pre-training, *k*	Post-training, *k*	*P* value of the difference
Composition	0.46 (95%CI: 0.37 to 0.54), moderate	0.52 (95%CI: 0.44 to 0.61), moderate	0.32
Echogenicity	0.36 (95%CI: 0.29 to 0.44), fair	0.44 (95%CI: 0.37 to 0.52), moderate	0.30
Shape	0.09 (95%CI: 0.02 to 0.21), slight	0.67 (95%CI: 0.56 to 0.78), substantial	< 0.001
Margins	0.03 (95%CI: -0.14 to 0.08), slight	0.05 (95%CI: -0.05 to 0.15), slight	0.71
Echogenic Foci	0.28 (95%CI: 0.19 to 0.37), fair	0.45 (95%CI: 0.36 to 0.53), moderate	0.004
TI-RADS Level	0.14 (95%CI: 0.08 to 0.20), slight	0.36 (95%CI: 0.30 to 0.42), fair	< 0.001
Recommendations	0.36 (95%CI: 0.27 to 0.45), fair	0.50 (95%CI: 0.41 to 0.59), moderate	0.02

Table 2 Percentage reader agreement with the reference standard for sonographic features

Sonographic feature	RS	$R1_{pre}$	$R1_{post}$	$R2_{pre}$	$R2_{post}$	$R3_{pre}$	$R3_{post}$
Composition	*n*	*n* (%)					
Spongiform	4	0 (0)	1 (25)	1 (25)	1 (25)	3 (75)	4 (100)
Cystic or almost completely cystic	11	3 (27.3)	5 (45.5)	7 (63.6)	8 (72.7)	10(90.9)	10(90.9)
Mixed cystic and solid	12	9 (75)	6 (50)	5 (41.7)	7 (58.3)	5 (58.3)	6 (50)
Solid	27	26 (96.3)	26 (96.3)	25 (92.6)	26 (96.3)	18 (66.7)	19 (70.4)
Echogenicity							
Anechoic	11	3 (27.3)	5 (45.5)	5 (45.5)	5 (45.5)	9 (81.8)	8 (72.7)
Hyperechoic or isoechoic	27	23 (85.2)	23 (85.2)	19 (70.4)	21 (77.8)	19 (70.4)	20 (74.1)
Hypoechoic	12	2 (16.7)	4 (33.3)	9 (75)	8 (66.7)	4 (33.3)	4 (33.3)
Shape							
Wilder than tall	42	38 (90.5)	39 (92.9)	7 (16.7)	39 (92.9)	41 (97.6)	40 (95.2)
Taller than wide	8	7 (87.5)	7 (87.5)	7 (87.5)	7 (87.5)	6 (75)	4 (50)
Margins							
Smooth or ill defined	47	36 (76.6)	35 (74.5)	35 (74.5)	33 (70.2)	43 (91.5)	45 (95.7)
Lobulated or irregular	3	1 (33.3)	2 (66.7)	1 (33.3)	2 (66.7)	0 (0)	0 (0)
Echogenic foci							
None or large comet tail artifact	41	20 (48.8)	36 (87.8)	29 (70.7)	39 (95.1)	29 (70.7)	29 (70.7)
Macrocalcification	3	1 (33.3)	1 (33.3)	0 (0)	2 (66.7)	2 (66.7)	2 (66.7)
Punctate echogenic foci	6	5 (83.3)	4 (66.7)	2 (33.3)	5 (83.3)	3 (50)	3 (50)

RS: Reference standard; R1: Reader 1; R2: Reader 2; R3: Reader 3.

with 50 nodules per group. After evaluating each group, a post-group training session was carried out for the freshman. The study found that the inter-reader agreement improved with training. Chung *et al*[13] performed a study evaluating the impact of radiologist's experience on ACR TI-RADS. Six fellowship-trained radiologists were divided into two groups (experienced *vs* less experienced) with the experienced group having at least 20 years of post-fellowship experience each and the less experienced group having 1 year or less of post-fellowship experience each. The study found no significant differences for inter-reader agreement between experienced *vs* less experienced readers for *'TI-RADS level'* or *'ACR TI-RADS recommendations'*. The inter-reader agreement was moderate for both experienced and less experienced groups for '*TI-RADS level*' and moderate to substantial (experienced *vs* less experienced, respectively) for '*ACR TI-RADS recommendations*'. Seifert *et al*[14] evaluated the inter-reader agreement and efficacy of consensus reading for several thyroid imaging risk stratification systems including ACR TI-RADS. The study involved 4 experienced

Table 3 Percentage reader agreement with the reference standard for American College of Radiology Thyroid Imaging Reporting and Data System levels

ACR TI-RADS level	RS, *n*	R1 $_{pre}$, *n* (%)	R1 $_{post}$, *n* (%)	R2 $_{pre}$, *n* (%)	R2 $_{post}$, *n* (%)	R3 $_{pre}$, *n* (%)	R3 $_{post}$, *n* (%)
1	11	1 (9.1)	5 (45.5)	1 (9.1)	7 (63.6)	10 (90.9)	8 (72.7)
2	9	3 (33.3)	4 (44.4)	0 (0)	4 (44.4)	3 (33.3)	3 (33.3)
3	9	4 (44.4)	5 (55.5)	1 (11.1)	6 (66.7)	4 (44.4)	6 (66.7)
4	13	4 (30.8)	5 (38.5)	5 (38.5)	9 (69.2)	5 (38.5)	5 (38.5)
5	8	7 (87.5)	4 (50)	6 (75)	5 (62.5)	3 (37.5)	3 (37.5)

ACR TI-RADS: American College of Radiology Thyroid Imaging Reporting and Data System; RS: Reference standard; R1: Reader 1; R2: Reader 2; R3: Reader 3.

Table 4 Percentage reader agreement with the reference standard for American College of Radiology Thyroid Imaging Reporting and Data System recommendations

Recommendations	RS, *n*	R1 $_{pre}$, *n* (%)	R1 $_{post}$, *n* (%)	R2 $_{pre}$, *n* (%)	R2 $_{post}$, *n* (%)	R3 $_{pre}$, *n* (%)	R3 $_{post}$, *n* (%)
No follow up	25	13 (52)	17 (68)	10 (40)	19 (76)	21 (84)	22 (88)
Follow up	5	3 (60)	1 (20)	1 (20)	3 (60)	3 (60)	3 (60)
FNA	20	17 (85)	15 (75)	18 (90)	17 (85)	11 (55)	13 (65)

RS: Reference standard; R1: Reader 1; R2: Reader 2; R3: Reader 3; FNA: Fine needle aspiration.

specialist readers with more than 5 years of clinical experience each. The readers independently scored 40 thyroid image datasets in session 1 followed by a joint consensus read (C1). After this, the process was repeated with independent scoring of 40 new image datasets in session 2, followed by another consensus read (C2). For ACR TI-RADS, the study found a significantly higher inter-reader agreement for session 2 (k = 0.57, moderate) *vs* session 1 (k = 0.32, fair) [$P < 0.01$], indicating that the addition of a consensus read had an impact in improving the inter-reader agreement.

Our study also evaluated the inter-reader agreement of individual sonographic features including composition, echogenicity, shape, margins, and echogenic foci. Our findings showed a significant improvement in inter-reader agreement with training for features such as '*shape*' (k = 0.09, slight $_{pre\text{-}training}$ versus k = 0.67, substantial $_{post\text{-}training}$, $P <$ 0.001) and '*echogenic foci*' (k = 0.28, fair $_{pre\text{-}training}$ versus k = 0.45, moderate $_{post\text{-}training}$, P = 0.004) but not for the others. The features with the strongest inter-reader agreement in our study were '*shape*' (k = 0.67 $_{post\text{-}training}$, substantial) and 'composition' (k = 0.52 $_{post\text{-}training}$, moderate). Hoang *et al*[7] also found similar findings in their study with 'shape' (k = 0.61, substantial) and 'composition' (k = 0.58, moderate) having the strongest inter-reader agreement amongst the 5 principal sonographic features. The feature with the poorest inter-reader agreement in our study was margins (k = 0.05 $_{post\text{-}training}$, slight). Similarly, Hoang *et al* [7] also found that 'margins' had the poorest inter-reader agreement (k = 0.25, fair) in their study. The poor inter-reader agreement for 'margins' is not surprising as accurate assessment requires a thorough review of the entire cine clip, rather than review of the still images only. Margins may also be harder to interpret through ultrasound artifacts. Finally, two of the available answer options for 'margins' in ACR TI-RADS are 'ill defined' (TI-RADS + 0 points) and 'irregular' (TI-RADS + 2 points). However, both options share innate conceptual similarities in interpretation and can lead to overlap. The poorest and strongest inter-reader agreement were also matched with the same features identified by Hoang's board-certified radiologists, indicating that the limitation may be inherent to the reporting and data system rather than trainee experience.

We also evaluated the relative sensitivity and specificity of the radiology residents in assigning TI-RADS levels compared to consensus reference standard before and after training. There was a general trend towards improved pooled sensitivity with TI-RADS levels 1 to 4 for the post-training assessment while the pooled specificity was relatively high (76.6-96.8%) for all TI-RADS level. Overall findings suggest that a single didactic training session improves the detection of benign (TI-RADS 1-3) lesions while

Table 5 The relative sensitivity, specificity, positive predictive value, and negative predictive value per Thyroid Imaging Reporting and Data System Level on the pre-training assessment compared to the reference standard

Pre-training, Statistics	TI-RADS 1, %	TI-RADS 2, %	TI-RADS 3, %	TI-RADS 4, %	TI-RADS 5, %
Sensitivity					
R1	9.1 (0.2-41.3)	33.3 (7.5-70.1)	44.4 (13.7-78.8)	30.8 (9.1-61.4)	87.5 (47.4-99.7)
R2	9.1 (0.2-41.3)	0 (0-33.6)	11.1 (0.3-48.3)	38.5 (13.9-68.4)	75 (34.9-96.8)
R3	90.9 (58.7-99.8)	33.3 (7.5-70.1)	44.4 (13.7-78.8)	38.5 (13.9-68.4)	37.5 (8.5-75.5)
Pooled	36.4 (20.4-54.9)	22.2 (8.6-42.3)	33.3 (16.5-54)	35.9 (21.2-52.8)	66.7 (44.7-84.4)
Specificity					
R1	100 (91.0-100)	90.2 (76.9-97.3)	92.7 (80.1-98.5)	62.2 (44.8-77.5)	76.2 (60.6-88)
R2	100 (91-100)	97.6 (87.1-99.9)	80.5 (65.1-91.2)	81.1 (64.8-92)	50 (34.2-65.8)
R3	66.7 (49.8-80.9)	97.6 (87.1-99.9)	95.1 (83.5-99.4)	89.2 (74.6-97)	90.5 (77.4-97.3)
Pooled	88.9 (81.8-94)	95.1 (89.7-98.2)	89.4 (82.6-94.3)	76.6 (67.6-84.1)	72.2 (63.5-79.8)
Positive predictive value					
R1	100	42.9 (16.8-73.6)	57.1 (26.4-83.2)	22.2 (10.3-41.6)	41.2 (27.7-56.1)
R2	100	0	11.1 (1.8-46.8)	41.7 (21.5-65.1)	22.2 (14.8-32.1)
R3	43.5 (32.2-55.5)	75 (26-96.2)	66.7 (30.1-90.3)	55.6 (28.3-79.8)	42.9 (17.1-73.2)
Pooled	48 (31.8-64.6)	50 (25.9-74.1)	40.9 (24.8-59.2)	35 (23.9-48)	31.4 (23.5-40.5)
Negative predictive value					
R1	79.6 (76.4-82.5)	86.1 (79.4-90.8)	88.4 (80.8-93.2)	71.9 (62.2-79.9)	97 (83.5-99.5)
R2	79.6 (76.4-82.5)	81.6 (80.9-82.4)	80.5 (75.8-84.5)	79 (70.4-85.6)	91.3 (75.3-97.3)
R3	96.3 (79.8-99.4)	87 (80.7-91.4)	88.6 (81.2-93.4)	80.5 (72.6-86.5)	88.4 (81.5-92.9)
Pooled	83.2 (79.2-86.6)	84.8 (81.9-87.3)	85.9 (82.3-88.9)	77.3 (72.5-81.5)	91.9 (86.5-95.3)

TI-RADS: Thyroid Imaging Reporting and Data System; RS: Reference standard; R1: Reader 1; R2: Reader 2; R3: Reader 3.

retaining high specificity in radiology residents. Improved identification of benign lesions is critical in avoiding unnecessary biopsies and interventions, a major aim of the ACR TI-RADS system.

The current study has several limitations. One limitation is the lack of a pathological reference standard. The reference standard was an expert consensus review by 3 board certified radiologists with Body Imaging fellowship and 1-14 years of clinical experience. However, it should be noted that this study is designed primarily to evaluate inter-reader reliability of radiology residents, and not the inherent performance of the ACR TI-RADS itself. As such, an expert consensus panel was deemed a practical reference standard, and one that simulates 'real world' clinical practice[9]. Another limitation is the relatively small number of cases used. However, even with this limited number of cases, we were able to show statistically significant improvements in inter-reader agreement for the two major outcome variables (TI-RADS level and ACR TI-RADS recommendations). While there is a relatively even distribution of TI-RADS levels among the test cases *via* non-random selection, there is uneven distribution of individual ultrasound features within the group. Of the 50 test cases, only 3 nodules demonstrated 'lobulated or irregular' margins (TI-RADS points +2), while the remaining 47 are 'smooth' or 'ill-defined' (TI-RADS points +0). A larger sample size can improve this and lead to more representative analysis of individual ultrasound features. Finally, training retention over time was not evaluated in this study, with the post-training testing performed two weeks after didactic and training case review.

Table 6 The relative sensitivity, specificity, positive predictive value, and negative predictive value per Thyroid Imaging Reporting and Data System Level on the post-training assessment compared to the reference standard

Post-training, Statistics	TI-RADS 1, %	TI-RADS 2, %	TI-RADS 3, %	TI-RADS 4, %	TI-RADS 5, %
Sensitivity					
R1	45.5 (16.8-76.6)	44.4 (13.7-78.8)	55.6 (21.2-86.3)	38.5 (13.9-68.4)	50 (15.7-84.3)
R2	63.6 (30.8-89.1)	44.4 (13.7-78.8)	66.7 (29.9-92.5)	69.2 (38.6-90.9)	62.5 (24.5-91.5)
R3	72.7 (39-94)	33.3 (7.5-70.1)	66.7 (29.9-92.5)	38.5 (13.9-68.4)	37.5 (8.5-75.5)
Pooled	60.6 (42.1-77.1)	40.7 (22.4-61.2)	63 (42.4-80.6)	48.7 (32.4-65.2)	50 (29.1-70.9)
Specificity					
R1	92.3 (79.1-98.4)	97.6 (87.1-99.9)	90.2 (76.9-97.3)	70.3 (53-84.1)	81 (65.9-91.4)
R2	94.9 (82.7-99.4)	97.6 (87.1-99.9)	95.1 (83.5-99.4)	73 (38.6-90.9)	90.5 (77.4-97.3)
R3	66.7 (49.8-80.9)	95.1 (83.5-99.4)	97.6 (87.1-99.9)	86.5 (71.2-95.5)	90.5 (77.4-97.3)
Pooled	84.6 (76.8-90.6)	96.8 (91.9-99.1)	94.3 (88.6-97.7)	76.6 (67.6-84.1)	87.3 (80.2-92.6)
Positive predictive value					
R1	62.5 (32-85.5)	80 (33.6-96.9)	55.6 (29.4-79)	31.3 (16.3-51.5)	33.3 (16.5-56)
R2	77.8 (45.8-93.6)	80 (33.6-96.9)	75 (41.8-92.6)	47.4 (32.2-63.1)	55.6 (29.9-78.6)
R3	38.1 (25.8-52.2)	60 (22.6-88.5)	85.7 (45.1-97.8)	50 (25.6-74.4)	42.9 (17.1-73.2)
Pooled	52.6 (40.1-64.8)	73.3 (48.6-88.9)	70.8 (52.8-84.1)	42.2 (31.5-53.8)	42.9 (29-57.9)
Negative predictive value					
R1	85.7 (77.6-91.2)	88.9 (81.7-93.5)	90.2 (81.6-95.1)	76.5 (66.8-84)	89.5 (80.7-94.5)
R2	90.2 (80.8-95.3)	88.9 (81.7-93.5)	92.9 (83.7-97)	87.1 (74.5-94)	92.7 (83.7-96.9)
R3	89.7 (76.3-95.9)	86.7 (80.3-91.2)	93 (84.1-97.1)	80 (71.9-86.2)	88.4 (81.5-92.9)
Pooled	88.4 (83.2-92.1)	88.2 (84.5-91.1)	92.1 (87.6-95)	81 (75.5-85.4)	90.2 (85.9-93.2)

TI-RADS: Thyroid Imaging Reporting and Data System; RS: Reference standard; R1: Reader 1; R2: Reader 2; R3: Reader 3.

CONCLUSION

Overall, the current study demonstrates a statistically significant improvement in inter-reader agreement among radiology residents, with no prior ACR TI-RADS experience, in the assignment of TI-RADS level and recommendations after a single didactic teaching session compared to expert consensus. Our study demonstrates the learnability of the ACR TI-RADS system and supports the use of dedicated training in radiology residents. Future studies can also be directed to evaluate the effect of additional training sessions with focus on areas/features demonstrating lower inter-rater agreement such as "margins" and retention of training over time.

ARTICLE HIGHLIGHTS

Research background

Thyroid nodules are common and often incidental. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) standardizes the use of ultrasound for thyroid nodule risk stratification.

Research motivation

Despite the widespread usage of this system, the learnability of TI-RADS has not been proven in radiology trainees.

Research objectives

To evaluate the inter-reader reliability amongst radiology trainees before and after TI-

RADS training.

Research methods

Three PGY-4 radiology residents were evaluated for inter-reader reliability with a 50 thyroid nodule data set before and after a 1-hour didactic teaching session and review of a training data set, with assessment performed 6 wk apart. Performance was compared to a consensus panel reference standard of three fellowship trained radiologists.

Research results

After one session of dedicated TI-RADS training, the radiology residents demonstrated statistically significant improvement in inter-reader agreement in subcategories of "shape", "echogenic foci", "TI-RADS level", and "recommendations" when compared with expert panel consensus. A trend towards higher pooled sensitivity for TI-RADS level 1-4 is also observed.

Research conclusions

Resident trainees demonstrated a statistically significant improvement in inter-reader agreement for both TI-RADS level and recommendations after training. This study demonstrates the learnability of the ACR TI-RADS.

Research perspectives

A multi-institutional and multi-national assessment of radiology resident diagnostic accuracy and inter-reader reliability of ACR TI-RADS classification and recommendations before and after training would improve the generalizability of these results.

REFERENCES

1 **Gharib H**, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L, Paschke R, Valcavi R, Vitti P; AACE/ACE/AME Task Force on Thyroid Nodules. American association of clinical endocrinologists, american college of endocrinology, and associazione medici endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules--2016 update. *Endocr Pract* 2016; **22**: 622-639 [PMID: 27167915 DOI: 10.4158/EP161208.GL]

2 **Grani G**, Lamartina L, Cantisani V, Maranghi M, Lucia P, Durante C. Interobserver agreement of various thyroid imaging reporting and data systems. *Endocr Connect* 2018; **7**: 1-7 [PMID: 29196301 DOI: 10.1530/EC-17-0336]

3 **Smith-Bindman R**, Lebda P, Feldstein VA, Sellami D, Goldstein RB, Brasic N, Jin C, Kornak J. Risk of thyroid cancer based on thyroid ultrasound imaging characteristics: results of a population-based study. *JAMA Intern Med* 2013; **173**: 1788-1796 [PMID: 23978950 DOI: 10.1001/jamainternmed.2013.9245]

4 **Lim H**, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in Thyroid Cancer Incidence and Mortality in the United States, 1974-2013. *JAMA* 2017; **317**: 1338-1348 [PMID: 28362912 DOI: 10.1001/jama.2017.2719]

5 **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**: 587-595 [PMID: 28372962 DOI: 10.1016/j.jacr.2017.01.046]

6 **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 [PMID: 29465333 DOI: 10.1148/radiol.2018171074]

7 **Hoang JK**, Middleton WD, Farjat AE, Teefey SA, Abinanti N, Boschini FJ, Bronner AJ, Dahiya N, Hertzberg BS, Newman JR, Scanga D, Vogler RC, Tessler FN. Interobserver Variability of Sonographic Features Used in the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol* 2018; **211**: 162-167 [PMID: 29702015 DOI: 10.2214/AJR.17.19192]

8 **Teng D**, Fu P, Li W, Guo F, Wang H. Learnability and reproducibility of ACR Thyroid Imaging, Reporting and Data System (TI-RADS) in postgraduate freshmen. *Endocrine* 2020; **67**: 643-650 [PMID: 31919768 DOI: 10.1007/s12020-019-02161-y]

9 **Pi Y**, Wilson MP, Katlariwala P, Sam M, Ackerman T, Paskar L, Patel V, Low G. Diagnostic accuracy and inter-observer reliability of the O-RADS scoring system among staff radiologists in a North American academic clinical setting. *Abdom Radiol (NY)* 2021; **46**: 4967-4973 [PMID: 34185128 DOI: 10.1007/s00261-021-03193-7]

10 **McHugh ML**. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012; **22**: 276-282 [PMID: 23092060 DOI: 10.11613/BM.2012.031]

11 **Gwet KL**. Testing the Difference of Correlated Agreement Coefficients for Statistical Significance. *Educ Psychol Meas* 2016; **76**: 609-637 [PMID: 29795880 DOI: 10.1177/0013164415596420]

12 **Li W**, Wang Y, Wen J, Zhang L, Sun Y. Diagnostic Performance of American College of Radiology TI-RADS: A Systematic Review and Meta-Analysis. *AJR Am J Roentgenol* 2021; **216**: 38-47 [PMID: 32603229 DOI: 10.2214/AJR.19.22691]

13 **Chung R**, Rosenkrantz AB, Bennett GL, Dane B, Jacobs JE, Slywotzky C, Smereka PN, Tong A, Sheth S. Interreader Concordance of the TI-RADS: Impact of Radiologist Experience. *AJR Am J Roentgenol* 2020; **214**: 1152-1157 [PMID: 32097031 DOI: 10.2214/AJR.19.21913]

14 **Seifert P**, Görges R, Zimny M, Kreissl MC, Schenke S. Interobserver agreement and efficacy of consensus reading in Kwak-, EU-, and ACR-thyroid imaging recording and data systems and ATA guidelines for the ultrasound risk stratification of thyroid nodules. *Endocrine* 2020; **67**: 143-154 [PMID: 31741167 DOI: 10.1007/s12020-019-02134-1]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: bpgoffice@wjgnet.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com