Thyroid Nodules on Ultrasound in Children and Young Adults: Comparison of Diagnostic Performance of Radiologists' Impressions, ACR TI-RADS, and a Deep Learning Algorithm

Jichen Yang, BSE[1], Laura C. Page, MD[2], Lars Wagner, MD[3], Benjamin Wildman-Tobriner, MD[4], Logan Bisset, MD[4], Donald Frush, MD[4], Maciej A. Mazurowski, PhD[1,4,5]

Pediatric Imaging · Original Research

Keywords

deep learning, pediatric, thyroid cancer, thyroid nodules, Thyroid Imaging Reporting and Data System, TI-RADS, ultrasound

Submitted: Jul 3, 2022
Revision requested: Jul 18, 2022
Revision received: Sep 13, 2022
Accepted: Oct 10, 2022
First published online: Oct 19, 2022
Version of record: Jan 11, 2023

An electronic supplement is available online at doi.org/10.2214/AJR.22.28231.

Based on a presentation at the American Thyroid Association 2022 annual meeting, Montreal, QC, Canada.

J. Yang and L. C. Page contributed equally to this work.

B. Wildman-Tobriner is a consultant for See-Mode Technologies. The remaining authors declare that there are no disclosures relevant to the subject matter of this article.

Supported in part by Derfner Foundation Award to L. C. Page.

BACKGROUND. In current clinical practice, thyroid nodules in children are generally evaluated on the basis of radiologists' overall impressions of ultrasound images.

OBJECTIVE. The purpose of this article is to compare the diagnostic performance of radiologists' overall impression, the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS), and a deep learning algorithm in differentiating benign and malignant thyroid nodules on ultrasound in children and young adults.

METHODS. This retrospective study included 139 patients (median age 17.5 years; 119 female patients, 20 male patients) evaluated from January 1, 2004, to September 18, 2020, who were 21 years old and younger with a thyroid nodule on ultrasound with definitive pathologic results from fine-needle aspiration and/or surgical excision to serve as the reference standard. A single nodule per patient was selected, and one transverse and one longitudinal image each of the nodules were extracted for further evaluation. Three radiologists independently characterized nodules on the basis of their overall impression (benign vs malignant) and ACR TI-RADS. A previously developed deep learning algorithm determined for each nodule a likelihood of malignancy, which was used to derive a risk level. Sensitivities and specificities for malignancy were calculated. Agreement was assessed using Cohen kappa coefficients.

RESULTS. For radiologists' overall impression, sensitivity ranged from 32.1% to 75.0% (mean, 58.3%; 95% CI, 49.2–67.3%), and specificity ranged from 63.8% to 93.9% (mean, 79.9%; 95% CI, 73.8–85.7%). For ACR TI-RADS, sensitivity ranged from 82.1% to 87.5% (mean, 85.1%; 95% CI, 77.3–92.1%), and specificity ranged from 47.0% to 54.2% (mean, 50.6%; 95% CI, 41.4–59.8%). The deep learning algorithm had a sensitivity of 87.5% (95% CI, 78.3–95.5%) and specificity of 36.1% (95% CI, 25.6–46.8%). Interobserver agreement among pairwise combinations of readers, expressed as kappa, for overall impression was 0.227–0.472 and for ACR TI-RADS was 0.597–0.643.

CONCLUSION. Both ACR TI-RADS and the deep learning algorithm had higher sensitivity albeit lower specificity compared with overall impressions. The deep learning algorithm had similar sensitivity but lower specificity than ACR TI-RADS. Interobserver agreement was higher for ACR TI-RADS than for overall impressions.

CLINICAL IMPACT. ACR TI-RADS and the deep learning algorithm may serve as potential alternative strategies for guiding decisions to perform fine-needle aspiration of thyroid nodules in children.

Thyroid cancer incidence in children has increased over the last several decades [1, 2]. However, the ability to noninvasively differentiate benign and malignant thyroid nodules in children remains limited. Certain ultrasound features are associated with an increased risk of malignancy for thyroid nodules in children [3]. Nonetheless, ultrasound has moderate sensitivity and low PPV for malignancy in this age group [3–5]. A method for accurate thyroid nodule assessment in children is important given that thyroid nodules in children have a risk of malignancy of 22–26%, compared with 5–10% in adults [6], and a risk of distant metastases as high as 30%, compared with 5% in adults [7].

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

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

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

doi.org/10.2214/AJR.22.28231

AJR 2023; 220:408–417

ISSN-L 0361–803X/23/2203–408

[1]Department of Electrical and Computer Engineering, Edmund T. Pratt Jr. School of Engineering, Duke University, Box 90291, Durham, NC 27708. Address correspondence to J. Yang (jy168@duke.edu).

[2]Department of Pediatrics, Division of Pediatric Endocrinology and Diabetes, Duke University School of Medicine, Durham, NC.

[3]Department of Pediatrics, Division of Pediatric Hematology and Oncology, Duke University School of Medicine, Durham, NC.

[4]Department of Radiology, Duke University School of Medicine, Durham, NC.

[5]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC.

To standardize thyroid nodule assessment and reduce unnecessary fine-needle aspirations (FNAs), the American College of Radiology (ACR) created a Thyroid Imaging Reporting and Data System (TI-RADS) [8]. This system assigns points for thyroid nodule features in five categories, which are then summed to derive an overall assessment category [8]. The ACR TI-RADS category is combined with the nodule's maximum diameter to reach one of three recommendations: perform FNA of the nodule, follow the nodule by serial ultrasound examinations, or do not perform further follow-up of the nodule [8]. ACR TI-RADS has been extensively studied in adults, and a meta-analysis found a pooled sensitivity of 89% and a pooled specificity of 70% [9]. Several studies have investigated the performance of ACR TI-RADS for evaluating thyroid nodules in children [10–14]. Although the outcomes of these studies have varied, ACR TI-RADS is generally viewed as lacking adequate sensitivity to be applied in children [10, 11]. This finding is not surprising given that ACR TI-RADS intentionally limits the detection of some small thyroid cancers that may be clinically insignificant in older patients [8]. However, given children's projected lifespans and their increased likelihood of advanced disease at the time of thyroid cancer diagnosis compared with adults [6], it is difficult to determine which, if any, pediatric thyroid cancers are clinically insignificant. Owing to the lack of a standardized scoring system that has been validated in children, ultrasound examinations of thyroid nodules in children are generally interpreted on the basis of the radiologist's overall impression [15].

Artificial intelligence methods, including deep learning, have been applied to the evaluation of thyroid nodules in adults [16–18]. In deep learning, networks of interconnected units identify patterns in data and subsequently use these patterns to perform complex tasks, for example, determining a thyroid nodule's likelihood of malignancy solely on the basis of its ultrasound features [19]. Whereas traditional machine learning methods (e.g., radiomics-based machine learning) rely on humans to define the characteristics to extract, deep learning convolutional neural networks identify important image features without human input and learn to perform classification during the training process. Thus, deep learning can incorporate characteristics that may not be recognizable by humans. To our knowledge, no previous studies have applied deep learning to the assessment of thyroid nodules in children. The aim of this study was to compare the diagnostic performance of radiologists' overall impressions, ACR TI-RADS, and a deep learning algorithm in differentiating benign and malignant thyroid nodules on ultrasound in children and young adults.

HIGHLIGHTS

Key Finding

- *For evaluation of thyroid nodules on ultrasound in children and young adults, radiologists' overall impression had mean sensitivity of 58.3% and mean specificity of 79.9%; ACR TI-RADS had mean sensitivity of 85.1% and mean specificity of 50.6%, and a deep learning algorithm had sensitivity of 87.5% and specificity of 36.1%.*

Importance

- *Both ACR TI-RADS and the deep learning algorithm warrant further exploration given the heightened priority for high sensitivity for evaluation of thyroid nodules in children.*

Methods

Patients

This retrospective study was approved by Duke University School of Medicine's institutional review board and was HIPAA compliant. The requirement for written informed consent was waived. The study was performed at a tertiary referral center. The institutional electronic medical record (EMR) was searched for consecutive patients who were 21 years or younger at the time of the encounter who had both an International Classification of Diseases (ICD) code consistent with a thyroid nodule or thyroid cancer and a pathology report containing the word "thyroid" between January 1, 2004, through September 18, 2020 (September 18 was the date the search was performed). This search used the following codes: ICD-9 codes 193 (malignant neoplasm of thyroid gland), 226 (benign neoplasm of thyroid gland), 242.40 (thyrotoxicosis from ectopic thyroid nodule), 242.41 (thyrotoxicosis from ectopic thyroid nodule with thyrotoxic crisis or storm), 244.0 (postsurgical hypothyroidism), 246.2 (cyst of thyroid), and V10.87 (personal history of malignant neoplasm of thyroid); and ICD-10 codes C73 (malignant neoplasm of thyroid gland), D34 (benign neoplasm of thyroid gland), D44.0 (neoplasm of uncertain behavior of thyroid gland), E04.1 (nontoxic single thyroid nodule), E05.10 (thyrotoxicosis with toxic single thyroid nodule without thyrotoxic crisis or storm), E05.11 (thyrotoxicosis with toxic single thyroid nodule with thyrotoxic crisis or storm), and Z85.850 (personal history of malignant neoplasm of thyroid). This search yielded 291 patients. A single investigator (L.C.P., a board-certified pediatric endocrinologist with 2 years of postfellowship experience) reviewed the search results, correlating nodules between ultrasound images and available pathologic results from FNA and/or surgical excision. On the basis of this review, 152 patients were excluded for the following reasons: no thyroid pathology report available (n = 8), diagnosis of multiple endocrine neoplasia type 2 (n = 11), no available thyroid ultrasound examination (n = 91), no thyroid nodule detected on ultrasound (n = 17), no thyroid nodule on ultrasound with definitive pathologic results based on FNA and/or surgical excision (n = 24), and age greater than 21 at the time of ultrasound despite being 21 years old or younger in initial EMR search results (n = 1). In the eight patients excluded because of lack of an available thyroid pathology report, the pathology report captured by the initial search was of a cervical lymph node. These exclusions resulted in a final study sample of 139 patients (median age, 17.5 years [IQR, 15.3–19.3 years]; 119 female patients, 20 male patients). Patients were not excluded on the basis of prior radiation treatment of cancer or on the basis of thyroid ultrasound having been performed at an outside institution (provided that the images were available for review). In each patient, the largest nodule with definitive pathologic results from FNA and/or surgical excision was selected for analysis. Figure 1 shows a flowchart of the patient selection process.

Preparation of Ultrasound Images

Ultrasound examinations were performed using a variety of ultrasound systems and scan techniques, reflecting the long dura-

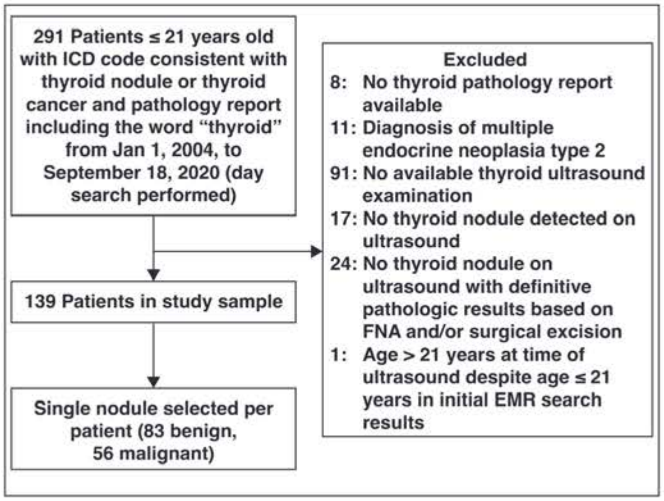
291 Patients ≤ 21 years old with ICD code consistent with thyroid nodule or thyroid cancer and pathology report including the word "thyroid" from Jan 1, 2004, to September 18, 2020 (day search performed)
Excluded
8: No thyroid pathology report available
11: Diagnosis of multiple endocrine neoplasia type 2
91: No available thyroid ultrasound examination
17: No thyroid nodule detected on ultrasound
24: No thyroid nodule on ultrasound with definitive pathologic results based on FNA and/or surgical excision
1: Age > 21 years at time of ultrasound despite age ≤ 21 years in initial EMR search results
139 Patients in study sample
Single nodule selected per patient (83 benign, 56 malignant)

Fig. 1—Flowchart shows patient selection process. ICD = International Classification of Diseases, FNA = fine-needle aspiration, EMR = electronic medical record.

tion of the study period and the inclusion of examinations performed at outside institutions. One investigator (J.Y., a graduate student) prepared ultrasound images of the single selected thyroid nodule per patient for subsequent analysis. The investigator selected a single static gray-scale transverse image and a single static gray-scale longitudinal image showing the nodule. Both selected images contained the calipers measuring the nodule that were placed by the technologist at the time of image acquisition. The investigator deidentified the images by removing the examination date and the patient's age and sex. Technical parameters such as frequency and depth were not removed from the images. The deidentified images were saved in PNG format. No additional image preprocessing was performed. Additional images of the thyroid gland, lymph nodes, and neck soft tissues were not evaluated as part of this investigation.

Interpretation of Ultrasound Images

Ultrasound examinations were independently interpreted by three radiologists (D.F., a fellowship-trained pediatric radiologist with 30 years of posttraining experience; L.B., a fellowship-trained pediatric radiologist with 1 year of posttraining experience; and B.W.T., an abdominal imaging fellowship-trained radiologist with 3 years of posttraining experience). The radiologists were informed that ultrasound examinations had been performed in patients 21 years or younger but were not informed of pathologic results or other clinical details. The radiologists were instructed to first categorize the depicted nodule as benign or malignant on the basis of their overall impression. They were instructed to then evaluate the nodule using ACR TI-RADS. For this purpose, the readers assessed nodule composition (classified as cystic or almost cystic, spongiform, mixed cystic and solid, solid or almost completely solid, or cannot be determined because of calcification), echogenicity (classified as anechoic, hyperechoic or isoechoic, hypoechoic, very hypoechoic, or cannot be determined), shape (classified as wider-than-tall or taller-than-wide), margin (classified as smooth, ill-defined, lobulated or irregular, extrathyroidal extension, or cannot be determined), no echogenic foci or large comet-tail artifacts (classified as criterion satisfied or not satisfied), macrocalcifications (classified as present or absent), peripheral calcifications (classified as present or absent), and punctate echogenic foci (classified as present or absent). The reader assessments for these features were used to automatically assign points for each ACR TI-RADS category (composition, echogenicity, shape, margin, and echogenic foci) without further reader input. One of the previously noted investigators who was not involved in blinded image review (L.C.P.) then used these points for each category to derive a total number of points and an overall ACR TI-RADS assessment category. Finally, the categories were used in combination with maximal nodule diameter (as extracted from the caliper measurements on the images) to derive the recommended management for the nodule on the basis of ACR TI-RADS (FNA, ultrasound follow-up, or no follow-up); the readers did not directly assign these management recommendations. Before the image evaluations, the readers were provided with written instructions on the use of ACR TI-RADS. In clinical practice, the two pediatric radiologists did not assign ACR TI-RADS scores or recommendations to thyroid nodules, whereas the abdominal imaging radiologist used ACR TI-RADS to assess approximately 20 ultrasound examinations per month.

Deep Learning Algorithm

The images evaluated by the three readers were also evaluated by a deep learning algorithm that was previously developed at Duke University School of Medicine [16]. This algorithm was created by training a multitask deep convolutional neural network from random weights (six convolutional layers and five maximum-pooling layers) on 1278 thyroid nodules from 1139 adults [16]. The algorithm derives a probability of malignancy for each nodule ranging from 0 to 1, which the algorithm then classifies into risk levels ranging from DL2 through DL5. These risk levels are distinct from the ACR TI-RADS categories TR1 through TR5. The algorithm does not incorporate a category of DL1. The algorithm also yields a management recommendation (FNA, ultrasound follow-up, or no follow-up) according to the risk level (DL2 through DL5) and nodule size (as entered by the user) [16]. The algorithm was previously validated on a test set of 99 thyroid nodules in adult patients (mean age, 53.2 years; range, 19–82 years) from a different institution from the current investigation [16]; no patients overlapped between the prior and current studies. For the current study, the algorithm was implemented on a Linux system computer with Python code.

Reference Standard

As part of patient care, cytopathology reports from FNA and from surgical specimens were reviewed by the institution's pathologists using the Bethesda System for Reporting Thyroid Cytopathology [20]. Nodules were considered to be definitively benign by cytopathology if they were categorized as Bethesda class II [20]. For purposes of this investigation, the pathology reports from FNA and surgical excision were reviewed; the original slides were not evaluated. The specific diagnosis was recorded for each nodule, except for nodules with definitively benign cytopathologic findings on FNA but without subsequent surgical excision. On the basis of the recorded diagnoses, nodules were classified as benign or malignant.

TABLE 1: Patient and Nodule Characteristics

Parameter	Value
Age (y)	
Median	17.5
IQR	15.3–19.3
Sex	
Female	85.6 (119/139)
Male	14.4 (20/139)
Nodule size	
Median (cm)	2.4
IQR (cm)	1.6–3.7
< 0.5 cm	0.0 (0/139)
0.5 to < 1.0 cm	5.0 (7/139)
1.0 to < 1.5 cm	15.8 (22/139)
1.5 to < 2.5 cm	30.9 (43/139)
≥ 2.5 cm	48.2 (67/139)
Available pathology	
Cytopathology only	24.5 (34/139)
Surgical pathology only	10.8 (15/139)
Both cytopathology and surgical pathology	64.7 (90/139)
Cytopathologic results	
Indeterminate	29.8 (37/124)
Bethesda III	19.4 (24/124)
Bethesda IV	10.5 (13/124)
Final classification	
Benign	59.7 (83/139)
Malignant	40.3 (56/139)
Papillary subtype (among malignant nodules)	89.3 (50/56)

Note—Unless otherwise indicated, values indicate percentages followed by numerator and denominator in parentheses. Numbers may not sum to 100 owing to rounding.

Statistical Analysis

Data were summarized using counts with percentages and medians with IQR. Sensitivity, specificity, NPV, and PPV for malignancy were calculated for the radiologists' overall impressions, ACR TI-RADS, and the deep learning algorithm. For purposes of determining diagnostic performance of ACR TI-RADS and the deep learning algorithm, nodules were considered benign if not recommended for FNA, such that malignant nodules recommended for ultrasound follow-up were considered to represent false-negative interpretations. Mean values across the three readers were obtained for the diagnostic performance measures for overall impression and ACR TI-RADS. The 95% CIs were estimated for the mean diagnostic performance measures for overall impression and ACR TI-RADS and for the diagnostic performance measures for the deep learning algorithm using bootstrapping with 10,000 repetitions [21]. Diagnostic performance was also evaluated in the subset of patients 18 years old and younger and compared qualitatively with results in the full study sample. Agreement in terms of classification as benign or malignant was assessed among all pairwise combinations of radiologists' overall impression, ACR TI-RADS, and the deep learning algorithm using Cohen kappa coefficient. Interobserver agreement among the three readers was also assessed for overall impression and for individual ACR TI-RADS features, overall ACR TI-RADS assessment category, and recommendation for FNA according to ACR TI-RADS using Cohen kappa coefficient for binary features and Fleiss kappa coefficient for categoric features. The three readers' agreement for total ACR TI-RADS points was assessed using the intraclass correlation coefficient (ICC). Agreement was assessed as follows [22]: less than 0.200, very weak; 0.200–0.399, weak; 0.400–0.599, moderate; 0.600–0.799, strong; 0.800 and greater, very strong. All statistical analyses were performed using SciPy (version 1.9.0.dev0+1653.e1d2f1c) [23].

Results

Patients

Table 1 summarizes characteristics of included patients and the single evaluated nodule per patient. The median nodule size on ultrasound was 2.4 cm (IQR, 1.6–3.7 cm; range, 0.5–6.2 cm). A total of 24.5% (34/139) of nodules were evaluated by cytopathology from FNA only, 10.8% (15/139) by surgical excision only, and 64.7% (90/139) by both cytopathology from FNA and by surgical excision. Thus, cytopathology was available for 89.2% (124/139) of nodules, whereas surgical excision was available for 75.5% (105/139) of nodules. Of the nodules with available cytopathology, the cytopathologic results were indeterminate in 29.8% (37/124). A total of 37 nodules both had definitive cytopathology and underwent subsequent surgical excision; 36 of these nodules had concordant results between the two methods in terms of benignity versus malignancy. The one discordant nodule was classified as Bethesda II on cytopathology but was determined to be papillary thyroid cancer arising in an adenomatoid nodule at excision; this nodule was classified as malignant for purposes of analysis. A total of 59.7% (83/139) of nodules were benign, and 40.3% (56/139) were malignant. Table S1 (available in the online supplement) summarizes nodules' pathologic diagnoses. Of the benign nodules, 41.0% (34/83) had definitively benign cytopathologic results from FNA without subsequent surgical excision, 30.1% (25/83) were follicular adenomas, 16.9% (14/83) were hyperplastic nodules, and the remainder comprised a spectrum of additional benign diagnoses. Of the malignant nodules, 89.3% (50/56) were papillary thyroid cancer and the remaining were follicular thyroid carcinoma (*n* = 5) or Hürthle cell carcinoma (*n* = 1). No nodule was a noninvasive follicular thyroid neoplasm with papillarylike nuclear features.

Diagnostic Performance

According to overall impression, the three readers classified 38.1% (53/139), 16.5% (23/139), and 51.8% (72/139) of nodules as malignant. Table 2 summarizes the distribution of category assignments by ACR TI-RADS and the deep learning algorithm. On the basis of ACR TI-RADS, the three readers classified 5.8–23.0% of nodules as TR1, 1.4–10.8% as TR2, 10.1–23.0% as TR3, 23.0–36.7% as TR4, and 23.7–35.3% as TR5. The deep learning algorithm classified 7.9% of nodules as DL2, 14.4% as DL3, 52.5% as DL4, and 25.2% as DL5.

Table 3 summarizes diagnostic performance for radiologists' overall impression, ACR TI-RADS, and deep learning. For overall

TABLE 2: Distribution of Category Assignments for ACR TI-RADS and Deep Learning Algorithm

Interpretation Method, Reader	Category 1	Category 2	Category 3	Category 4	Category 5
ACR TI-RADS					
Reader 1	23.0 (32/139)	8.6 (12/139)	10.1 (14/139)	23.0 (32/139)	35.3 (49/139)
Reader 2	5.8 (8/139)	10.8 (15/139)	23.0 (32/139)	36.7 (51/139)	23.7 (33/139)
Reader 3	22.3 (31/139)	1.4 (2/139)	15.8 (22/139)	34.5 (48/139)	25.9 (36/139)
Deep learning algorithm		7.9 (11/139)	14.4 (20/139)	52.5 (73/139)	25.2 (35/139)

Note—Data represent percentages with numerator and denominator in parentheses. Categories correspond with TR1 through TR5 for ACR TI-RADS and for DL2 through DL5 for deep learning algorithm; deep learning algorithm does not incorporate a category of DL1. ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System.

TABLE 3: Diagnostic Performance of Radiologists' Overall Impression, ACR TI-RADS, and Deep Learning Algorithm

Interpretation Method, Reader	Sensitivity	Specificity	NPV	PPV
Overall impression				
Reader 1	67.9 (38/56)	81.9 (68/83)	79.1 (68/86)	71.7 (38/53)
Reader 2	32.1 (18/56)	94.0 (78/83)	67.2 (78/116)	78.3 (18/23)
Reader 3	75.0 (42/56)	63.9 (53/83)	79.1 (53/67)	57.5 (42/73)
Mean	58.3	79.9	75.1	69.2
95% CI	49.2–67.3	73.8–85.7	67.0–82.5	58.1–79.7
ACR TI-RADS				
Reader 1	85.7 (48/56)	50.6 (42/83)	84.0 (42/50)	53.9 (48/89)
Reader 2	82.1 (46/56)	47.0 (39/83)	79.6 (39/49)	51.1 (46/90)
Reader 3	87.5 (49/56)	54.2 (45/83)	86.5 (45/52)	56.3 (49/87)
Mean	85.1	50.6	83.4	53.7
95% CI	77.3–92.1	41.4–59.8	73.9–91.5	44.0–63.7
Deep learning algorithm				
Mean	87.5 (49/56)	36.1 (30/83)	81.1 (30/37)	48.0 (49/102)
95% CI	78.3–95.5	25.6–46.8	67.5–92.9	38.6–57.8

Note—Unless otherwise stated, data represent percentages and numerator and denominator in parentheses. ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System.

impression, sensitivity ranged from 32.1% to 75.0% (mean, 58.3%; 95% CI, 49.2–67.3%), and specificity ranged from 63.8% to 93.9% (mean, 79.9%; 95% CI, 73.8–85.7%). For ACR TI-RADS, sensitivity ranged from 82.1% to 87.5% (mean, 85.1%; 95% CI, 77.3–92.1%), and specificity ranged from 47.0% to 54.2% (mean, 50.6%; 95% CI, 41.4–59.8%). For each radiologist, sensitivity was higher and specificity was lower for ACR TI-RADS than for overall impression. The deep learning algorithm had a sensitivity of 87.5% (95% CI, 78.3–95.5%) and a specificity of 36.1% (95% CI, 25.6–46.8%). Table S2 (available in the online supplement) presents corresponding diagnostic performance information among a subset of 77 patients 18 years old or younger, showing qualitatively similar results; sensitivity and specificity were within 5% between the entire study sample and the subset in terms of mean performance of overall impression, mean performance of ACR TI-RADS, and the deep learning algorithm.

Figure 2 shows representative benign and malignant nodules, along with the radiologists' overall impressions and the management recommendations according to ACR TI-RADS and the deep learning algorithm.

Intermethod and Interobserver Agreement

Table 4 shows pairwise assessments of intermethod and interobserver agreement of nodule categorization as benign or malignant between radiologists' overall impression, ACR TI-RADS category, and the deep learning algorithm. Agreement, expressed as kappa, for pairwise combinations of overall impression among radiologists was weak to moderate (κ = 0.227–0.472), of ACR TI-RADS category among radiologists was moderate to strong (κ = 0.597–0.643), of overall impression and ACR TI-RADS category was very weak to moderate (κ = 0.171–0.487), of overall impression and the deep learning algorithm was very weak

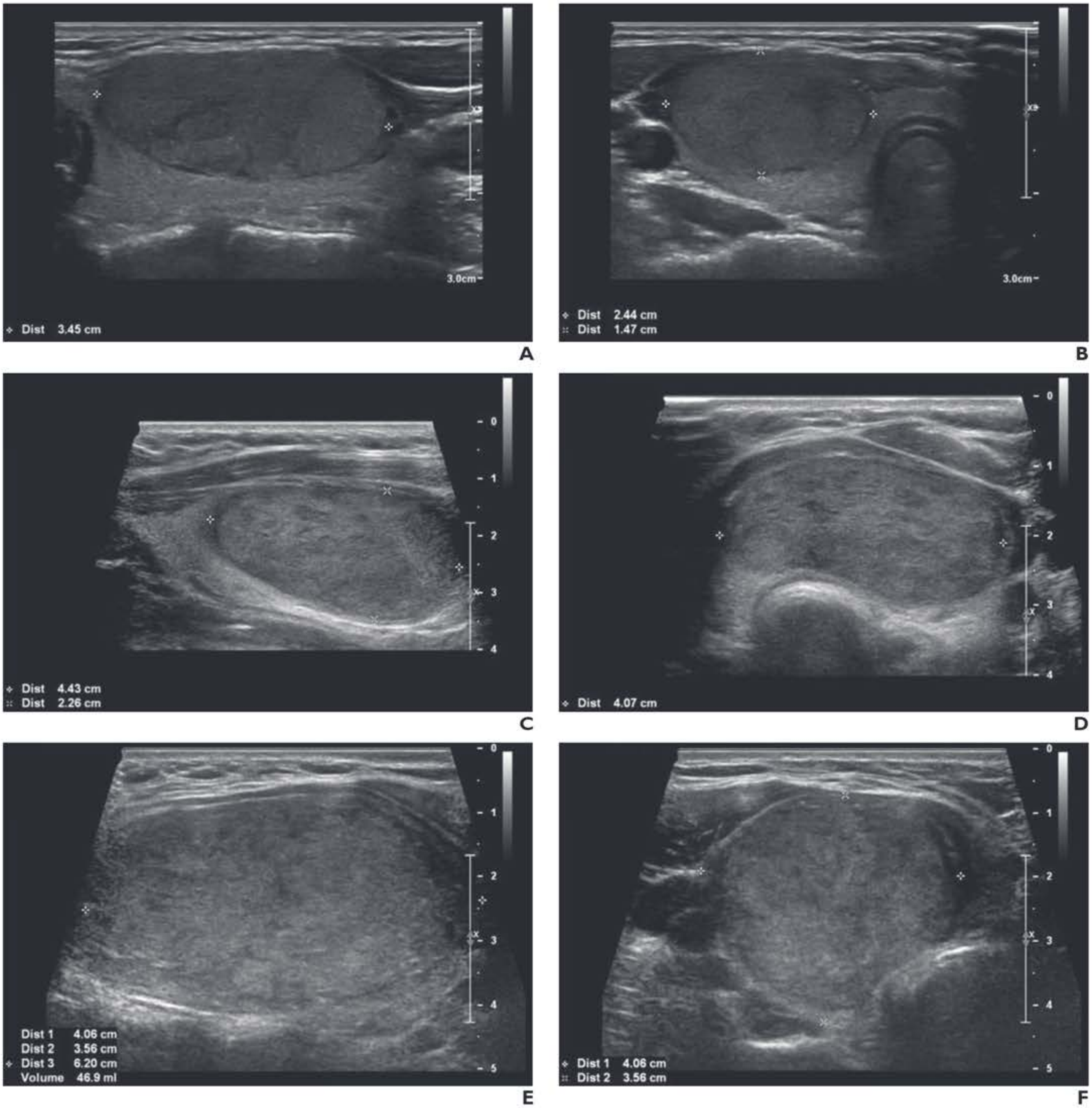

Fig. 2—Gray-scale ultrasound images of thyroid nodule in five patients, along with assessment by radiologists' overall impression, American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS), and deep learning algorithm. Dist = distance between calipers, FNA = fine-needle aspiration.

A and B, Longitudinal (**A**) and transverse (**B**) gray-scale ultrasound images show right lobe in 17-year-old patient with benign follicular adenoma according to surgical excision alone. Overall impression for all three radiologists was benign (true-negative). ACR TI-RADS recommendation for all three radiologists was FNA (false-positive). Deep learning algorithm recommendation was FNA (false-positive).

C and D, Longitudinal (**C**) and transverse (**D**) gray-scale ultrasound images show left lobe in 19-year-old patient with benign follicular adenoma according to surgical excision alone. Overall impression for radiologists 1 and 2 was benign (true-negative), and for radiologist 3 was malignant (false-positive). ACR TI-RADS recommendation for all radiologists was FNA (false-positive). Deep learning algorithm recommendation was FNA (false-positive).

E and F, Longitudinal (**E**) and transverse (**F**) gray-scale ultrasound images show right lobe in 20-year-old patient with papillary thyroid carcinoma according to cytopathology and surgical excision. Overall impression for radiologist 1 was malignant (true-positive) and for radiologists 2 and 3 was benign (false-negative). ACR TI-RADS recommendations for all three radiologists was FNA (true-positive). Deep learning algorithm recommendation was FNA (true-positive).

(Fig. 2 continues on next page)

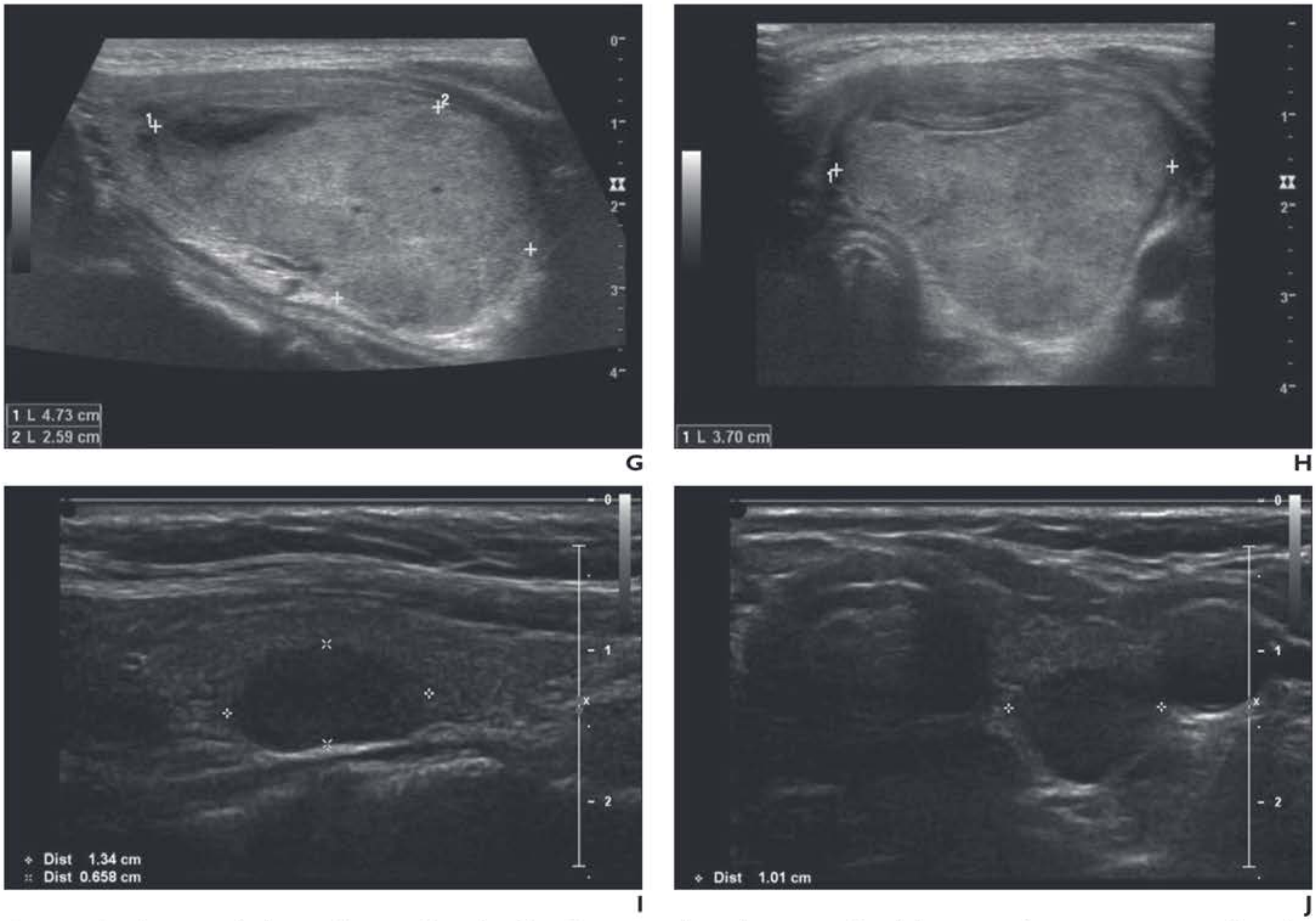

Fig. 2 (continued)—Gray-scale ultrasound images of thyroid nodule in five patients, along with assessment by radiologists' overall impression, American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS), and deep learning algorithm. Dist = distance between calipers, FNA = fine-needle aspiration.
G and **H**, Longitudinal (**G**) and transverse (**H**) gray-scale ultrasound images show left lobe in 18-year-old patient with follicular thyroid carcinoma according to surgical excision alone. Overall impression for all three radiologists was benign (false-negative). ACR TI-RADS recommendation for all three radiologists was FNA (true-positive). Deep learning recommendation was FNA (true-positive). L = length.
I and **J**, Longitudinal (**I**) and transverse (**J**) gray-scale ultrasound images show left lobe in 13-year-old patient with benign follicular adenoma according to surgical excision alone. Overall impression for radiologists 1 and 2 was benign (true-negative) and for radiologist 3 was malignant (false-positive). ACR TI-RADS recommendation for radiologist 1 was no further follow-up (true-negative) and for radiologists 2 and 3 was ultrasound follow-up (true-negative). Deep learning algorithm recommendation was no further follow-up (true-negative).

(κ = 0.063–0.184), and of ACR TI-RADS category and the deep learning algorithm was moderate (κ = 0.494–0.553).

Table 5 summarizes interobserver agreement among the three readers. Agreement was weak for overall impression (κ = 0.340). In terms of ACR TI-RADS features, agreement was strong for composition (κ = 0.650), moderate for echogenicity and shape (κ = 0.480–0.502), and weak for margin (κ = 0.203). In terms of features relating to echogenic foci, interobserver agreement was strong for absence of echogenic foci or large comet-tail artifacts (κ = 0.693), moderate for peripheral calcifications (κ = 0.419) and punctate echogenic foci (κ = 0.582), and weak for macrocalcifications (κ = 0.170). Interobserver agreement was strong for total points (ICC = 0.633), moderate for ACR TI-RADS assessment category (κ = 0.403), and strong for recommendation for FNA according to ACR TI-RADS (κ = 0.616).

Discussion

In this study, we evaluated various methods for differentiating benign and malignant thyroid nodules in children and young adults. Radiologists' overall impression (mean of three readers) had a sensitivity of 58.3% and 79.9%. ACR TI-RADS (mean of three readers) had sensitivity of 85.1% and specificity of 50.6%. A deep learning algorithm had sensitivity of 87.5% and 36.1%. These results indicate higher sensitivity, albeit lower specificity, of ACR TI-RADS and the deep learning algorithm with respect to radiologists' overall impression, and similar sensitivity but lower specificity of the deep learning algorithm with respect to ACR TI-RADS. The findings indicate the potential role of alternate strategies for guiding decisions to perform FNA of thyroid nodules in children, in comparison with the current approach of basing such management primarily on radiologists' overall impressions.

TABLE 4: Agreement Between Pairwise Combinations of Radiologists' Overall Impression, ACR TI-RADS Category, and Deep Learning Algorithm

	Impression			ACR TI-RADS		
Method	Radiologist 1	Radiologist 2	Radiologist 3	Radiologist 1	Radiologist 2	Radiologist 3
Impression						
Radiologist 2	0.418 (0.272–0.560)					
Radiologist 3	0.472 (0.329–0.611)	0.227 (0.116–0.346)				
ACR TI-RADS						
Radiologist 1	0.487 (0.369–0.611)	0.176 (0.096–0.268)	0.288 (0.131–0.442)	.		
Radiologist 2	0.314 (0.183–0.449)	0.171 (0.091–0.262)	0.215 0.056–0.370)	0.608 (0.465–0.738)		
Radiologist 3	0.403 (0.273–0.534)	0.187 (0.103–0.282)	0.376 (0.224–0.522)	0.597 (0.452–0.729)	0.643 (0.501–0.767)	
Deep learning algorithm	0.184 0.066–0.305)	0.112 (0.051–0.185)	0.063 (−0.083 to 0.212)	0.553 (0.399–0.695)	0.499 (0.343–0.644)	0.494 (0.337– 0.641)

Note—Data expressed as Cohen kappa coefficient, with 95% CIs in parentheses. Cells are blank for comparisons of method with itself or for duplicate of comparison appearing elsewhere in table. ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System.

Previous studies of ACR TI-RADS in children compared with the current study found similar sensitivity (ranging from 78% to 100%) and somewhat higher specificity (ranging from 69% to 79%) [10, 12, 14]. In addition, a prior study in children evaluated the performance of radiologists' overall impression, determined both by the consensus of two readers and by a third independent reader; compared with the current study, that study yielded higher sensitivities (81.0% and 94.0%, respectively) and similar specificities (76.0% and 81.0%, respectively) [13]. To our knowledge, the current study is the first to directly compare the diagnostic performance in children of ACR TI-RADS categories and radiologists' overall impressions. This comparison is important given that the clinical application of ACR TI-RADS is not standardized in children as it is in adults. Prior studies of ACR TI-RADS alone described the system as inadequate for use in children given an unacceptably high rate of false-negatives [10, 13]. However, in the current study, all radiologists had higher sensitivity (albeit lower specificity) using ACR TI-RADS than the overall impression, and interobserver agreement was higher for ACR TI-RADS than for overall impression as well.

Human interpretation of ultrasound images and pathologic interpretation of cytopathology from FNA have limitations in definitively characterizing thyroid nodules, and additional diagnostic methods are needed. Deep learning has potential advantages

TABLE 5: Interobserver Agreement Among Three Radiologists

Outcome	Agreement	95% CI
Overall impression	0.340	0.223–0.456
Individual ACR TI-RADS features		
Composition	0.650	0.558–0.739
Echogenicity	0.502	0.408–0.588
Shape	0.480	0.326–0.614
Margin	0.203	0.117–0.289
No echogenic foci or large comet-tail artifacts	0.693	0.577–0.794
Macrocalcifications	0.170	−0.043 to 0.494
Peripheral calcifications	0.419	−0.017 to 0.898
Punctate echogenic foci	0.582	0.451–0.695
Total ACR TI-RADS points	0.633	0.553–0.713
ACR TI-RADS assessment category	0.403	0.317–0.482
Recommendation for FNA based on ACR TI-RADS	0.616	0.508–0.714

Note—Data expressed as Cohen kappa coefficient, Fleiss kappa coefficient, or intraclass correlation coefficient. ACR TI-RADS = American College of Radiology Thyroid Imaging Reporting and Data System, FNA = fine-needle aspiration.

over traditional methods given that it is objective and reproducible and does not require access to a physician with dedicated pediatric training. In adults, high specificity in the diagnostic evaluation of thyroid nodules is important given the desire to limit unnecessary biopsies. However, in children, a thyroid cancer that is small or has low aggressiveness has a prolonged period to grow and/or metastasize [10]. Thus, in children, high sensitivity is a priority. Given the deep learning algorithm's relatively high sensitivity, the algorithm may be useful to help identify potentially malignant nodules for further radiologist evaluation at institutions that are currently relying solely on radiologists' overall impression without the use of ACR TI-RADS. Nonetheless, although the deep learning algorithm had high sensitivity, this sensitivity was similar to that of ACR TI-RADS, and the algorithm had very low specificity. Thus, the results do not strongly support the use of deep learning at this time and indicate the need for additional steps, including further training and validation using thyroid nodules in children, before the tool is applied clinically.

This study had limitations. First, the deep learning algorithm was trained and the threshold probability of malignancy for classifying nodules from DL2 to DL5 was determined using thyroid nodules from adults. The deep learning algorithm previously showed in adults a similar sensitivity (87%) but somewhat higher specificity (52%) [16] in comparison with its observed performance in children in the current study. Because thyroid cancer in children has unique clinical and molecular characteristics [24], training the deep convolutional neural network in children would likely improve its performance. Given that thyroid cancer is uncommon in children and that many cases are needed to train a deep convolutional neural network, the development of a robust deep learning algorithm using only data from children would likely require a multiinstitutional collaboration. Second, the study sample included patients 21 years old and younger. The optimal age threshold for differentiating pediatric versus adult thyroid cancer remains controversial [25]. However, similar results were obtained in a subanalysis of patients 18 years old and younger. Third, the radiologists were instructed to provide their overall impression of whether each nodule was benign or malignant, rather than whether or not they would recommend that the nodule undergo FNA. The radiologists may have elected to recommend FNA for some nodules for which they had an overall benign impression because of diagnostic uncertainty. Fourth, the radiologists and the deep learning algorithm evaluated only two static gray-scale images for each nodule. Performance may have been higher if evaluating all available static and cine images. Fifth, the frequency of malignancy of 40.3% was high. The frequency of malignancy may reflect referral bias from the tertiary-care study setting and selection bias given the inclusion only of nodules with definitive pathologic results. An earlier study of thyroid nodules in children had a similar high malignancy rate of 42% [13]. Sixth, 89.2% of cancers were of the papillary subtype. The frequency of this subtype, although expected in children [1], limits the ability to assess the performance of the various methods for diagnosing other thyroid cancer subtypes. Finally, the ultrasound image features used by the deep learning algorithm to differentiate benign and malignant nodules remain unknown from this analysis.

In conclusion, when evaluating thyroid nodules in children and young adults, both ACR TI-RADS and a deep learning algorithm previously trained in adults had higher sensitivity albeit lower specificity compared with radiologists' overall impressions (representing the current standard clinical approach). The deep learning algorithm had similar sensitivity but lower specificity than ACR TI-RADS. Further training and validation using pediatric data will be required before potential clinical application of the deep learning algorithm. Nonetheless, given the heightened priority for sensitivity when evaluating thyroid nodules in children compared with in adults, the findings support the continued exploration in children of ACR TI-RADS and of the deep learning algorithm.

Acknowledgments

We thank our patients and their families and the Derfner Foundation for their generous contributions to the Department of Pediatrics at Duke University.

Provenance and review: Not solicited; externally peer reviewed.

Peer reviewers: Sheng-Yang Huang, Taichung Veterans General Hospital; Liliane H. Gibbs, University of California Irvine; Filiz Elbuken, Hospital of the University of Pennsylvania; additional individuals who chose not to disclose their identities.

References

1. Qian ZJ, Jin MC, Meister KD, Megwalu UC. Pediatric thyroid cancer incidence and mortality trends in the United States, 1973–2013. *JAMA Otolaryngol Head Neck Surg* 2019; 145:617–623
2. Vergamini LB, Frazier AL, Abrantes FL, Ribeiro KB, Rodriguez-Galindo C. Increase in the incidence of differentiated thyroid carcinoma in children, adolescents, and young adults: a population-based study. *J Pediatr* 2014; 164:1481–1485
3. Richman DM, Benson CB, Doubilet PM, et al. Thyroid nodules in pediatric patients: sonographic characteristics and likelihood of cancer. *Radiology* 2018; 288:591–599
4. Richman DM, Cherella CE, Smith JR, et al. Clinical utility of sonographic features in indeterminate pediatric thyroid nodules. *Eur J Endocrinol* 2021; 184:657–665
5. Al Nofal A, Gionfriddo MR, Javed A, et al. Accuracy of thyroid nodule sonography for the detection of thyroid cancer in children: systematic review and meta-analysis. *Clin Endocrinol (Oxf)* 2016; 84:423–430
6. Francis GL, Waguespack SG, Bauer AJ, et al.; American Thyroid Association Guidelines Task Force. Management guidelines for children with thyroid nodules and differentiated thyroid cancer. *Thyroid* 2015; 25:716–759
7. Banik GL, Shindo ML, Kraimer KL, et al. Prevalence and risk factors for multifocality in pediatric thyroid cancer. *JAMA Otolaryngol Head Neck Surg* 2021; 147:1100–1106
8. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol* 2017; 14:587–595
9. Li W, Wang Y, Wen J, Zhang L, Sun Y. Diagnostic performance of American College of Radiology TI-RADS: a systematic review and meta-analysis. *AJR* 2021; 216:38–47
10. Richman DM, Benson CB, Doubilet PM, et al. Assessment of American College of Radiology Thyroid Imaging Reporting and Data System (TI-RADS) for pediatric thyroid nodules. *Radiology* 2020; 294:415–420
11. Ahmad H, Al-Hadidi A, Bobbey A, et al. Pediatric adaptions are needed to improve the diagnostic accuracy of thyroid ultrasound using TI-RADS. *J Pediatr Surg* 2021; 56:1120–1125
12. Polat YD, Öztürk VS, Ersoz N, Anık A, Karaman CZ. Is Thyroid Imaging Reporting and Data System useful as an adult ultrasonographic malignancy

risk stratification method in pediatric thyroid nodules? *J Med Ultrasound* 2019; 27:141–145

13. Martinez-Rios C, Daneman A, Bajno L, van der Kaay DCM, Moineddin R, Wasserman JD. Utility of adult-based ultrasound malignancy risk stratifications in pediatric thyroid nodules. *Pediatr Radiol* 2018; 48:74–84
14. Shapira-Zaltsberg G, Miller E, Martinez-Rios C, et al. Comparison of the diagnostic performance of the 2017 ACR TI-RADS guideline to the Kwak guideline in children with thyroid nodules. *Pediatr Radiol* 2019; 49:862–868
15. Kim PH, Yoon HM, Hwang J, et al. Diagnostic performance of adult-based ATA and ACR-TIRADS ultrasound risk stratification systems in pediatric thyroid nodules: a systematic review and meta-analysis. *Eur Radiol* 2021; 31:7450–7463
16. Buda M, Wildman-Tobriner B, Hoang JK, et al. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. *Radiology* 2019; 292:695–701
17. Zhou H, Jin Y, Dai L, et al. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. *Eur J Radiol* 2020; 127:108992
18. Song J, Chai YJ, Masuoka H, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine (Baltimore)* 2019; 98:e15133
19. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* 2019; 49:939–954
20. Cibas ES, Ali SZ. The Bethesda system for reporting thyroid cytopathology. *Thyroid* 2009; 19:1159–1165
21. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1986; 1:54–75
22. Campbell MJ, Swinscow TDV, eds. *Statistics at square one*, 9th ed. BMJ Publishing Group, 1996:140
23. Virtanen P, Gommers R, Oliphant TE, et al.; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; 17:261–272
24. Nies M, Vassilopoulou-Sellin R, Bassett RL, et al. Distant metastases from childhood differentiated thyroid carcinoma: clinical course and mutational landscape. *J Clin Endocrinol Metab* 2021; 106:e1683–e1697
25. Sugino K, Nagahama M, Kitagawa W, et al. Cutoff age between pediatric and adult thyroid differentiated cancer: is 18 years old appropriate? *Thyroid* 2021; 32:145–152

Editorial Comment: ACR TI-RADS for Pediatric Thyroid Nodules on Ultrasound—A Structured Interpretation

Thyroid nodules have lower incidence in children than in adults; however, the malignancy rate in children is higher. Ultrasound is the most common tool to identify potentially malignant thyroid lesions aside from laboratory testing [1]. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) is the most widely used system for structured evaluation of thyroid nodules on ultrasound, providing an alternative to gestalt impressions. ACR TI-RADS has shown variable diagnostic performance in adolescents and younger children. Given the practical difficulties in performing fine-needle aspiration (FNA) of small thyroid nodules in children, ultrasounds' sensitivity and specificity are both important.

For radiologists who are not currently applying ACR TI-RADS for evaluation of thyroid nodules in children, this study provides solid evidence to adopt the five-category system for future reports. Use of ACR TI-RADS decreased false-negatives for malignancy compared with radiologists' impressions. The findings also suggest radiologists' use of ACR TI-RADS is not inferior to a deep learning algorithm, which is relevant given growing interest in artificial intelligence for image interpretation. An earlier meta-analysis found ACR TI-RADS in children with a nodule with category 4 or 5 assessment to have a pooled sensitivity and specificity for detecting malignancy of 0.84 and 0.64, respectively [2]; however, that study did not compare ACR TI-RADS with radiologists' impressions or with a deep learning algorithm.

A key strength of the study is the pathologic reference standard in all patients. In addition, the study's Figure 2 provides meaningful examples of the use of ACR TI-RADS in children and young adults. A main limitation is the lack of derivation of nodule size thresholds in children that merit FNA or more aggressive intervention; this issue will require consensus by pediatricians, pediatric radiologists, and pediatric surgeons. Overall, this study advances the body of literature supporting use of ACR TI-RADS in children.

Sheng-Yang Huang, MD
Taichung Veterans General Hospital
Taichung, Taiwan
drugholic@vghtc.gov.tw

Version of record: Jan 11, 2023

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

doi.org/10.2214/AJR.22.28664

Provenance and review: Solicited; not externally peer reviewed.

References

1. Goldfarb M, Dinauer C. Differences in the management of thyroid nodules in children and adolescents as compared to adults. *Curr Opin Endocrinol Diabetes Obes* 2022; 29:466–473
2. Kim PH, Yoon HM, Hwang J, et al. Diagnostic performance of adult-based ATA and ACR-TIRADS ultrasound risk stratification systems in pediatric thyroid nodules: a systematic review and meta-analysis. *Eur Radiol* 2021; 31:7450–7463