Computer-Aided Diagnostic System for Thyroid Nodules on Ultrasonography: Diagnostic Performance Based on the Thyroid Imaging Reporting and Data System Classification and Dichotomous Outcomes

M. Han, E.J. Ha, and J.H. Park

ABSTRACT

BACKGROUND AND PURPOSE: Artificial intelligence-based computer-aided diagnostic systems have been introduced for thyroid cancer diagnosis. Our aim was to compare the diagnostic performance of a commercially available computer-aided diagnostic system and radiologist-based assessment for the detection of thyroid cancer based on the Thyroid Imaging Reporting and Data Systems (TIRADS) and dichotomous outcomes.

MATERIALS AND METHODS: In total, 372 consecutive patients with 454 thyroid nodules were enrolled. The computer-aided diagnostic system was set up to render a possible diagnosis in 2 formats, the Korean Society of Thyroid Radiology (K)-TIRADS and the American Thyroid Association (ATA)-TIRADS-classifications, and dichotomous outcomes (possibly benign or possibly malignant).

RESULTS: The diagnostic sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of the computer-aided diagnostic system for thyroid cancer were, respectively, 97.6%, 21.6%, 42.0%, 93.9%, and 49.6% for K-TIRADS; 94.6%, 29.6%, 43.9%, 90.4%, and 53.5% for ATA-TIRADS; and 81.4%, 81.9%, 72.3%, 88.3%, and 81.7% for dichotomous outcomes. The sensitivities of the computer-aided diagnostic system did not differ significantly from those of the radiologist (all $P > .05$); the specificities and accuracies were significantly lower than those of the radiologist (all $P < .001$). Unnecessary fine-needle aspiration rates were lower for the dichotomous outcome characterizations, particularly for those performed by the radiologist. The interobserver agreement for the description of K-TIRADS and ATA-TIRADS classifications was fair-to-moderate, but the dichotomous outcomes were in substantial agreement.

CONCLUSIONS: The diagnostic performance of the computer-aided diagnostic system varies in terms of TIRADS classification and dichotomous outcomes and relative to radiologist-based assessments. Clinicians should know about the strengths and weaknesses associated with the diagnosis of thyroid cancer using computer-aided diagnostic systems.

ABBREVIATIONS: AI = artificial intelligence; ATA = American Thyroid Association; CAD = computer-aided diagnosis; FNA = fine-needle aspiration; K = Korean; TIRADS = Thyroid Imaging Reporting and Data System; US = ultrasonography

Artificial intelligence (AI)-based computer-aided diagnostic (CAD) systems are projected to substantially influence the field of diagnostic imaging.[1,2] In recent years, the application of AI, particularly deep-learning techniques using convolutional neural networks, has shown promising results in radiology. However, because AI technologic innovation does not guarantee its usefulness in real-world medical practice, rigorous external clinical validation is necessary to determine its utility.[1-3]

Several AI-based CAD systems have shown potential in the field of thyroid imaging.[4-7] However, most reports describe proof-of-concept technical feasibility studies and lack robust validation.[8] One AI-based CAD system has recently been integrated into a commercially available ultrasonography (US) platform for thyroid imaging: the S-Detect CAD system (Samsung Medison). The system generates 2 outputs: Thyroid Imaging Reporting and Data System (TIRADS)-based scoring and dichotomous predictions. The dichotomous prediction is a completely independent diagnosis based on convolutional neural network deep learning techniques. However, commercialized CAD systems have not yet undergone rigorous validation, and few articles have described the diagnostic performance using dichotomous outcomes or compared the sensitivity relative to radiologist-based assessments.[9-12] TIRADS classification has been widely

Received February 14, 2020; accepted after revision September 29.

From the Department of Radiology, Ajou University School of Medicine, Suwon, Korea.

This work was supported by the National Research Foundation of Korea (No. 2017R1C1B5016217).

Please address correspondence to Eun Ju Ha, MD, PhD, Department of Radiology, Ajou University School of Medicine, Wonchon-Dong, Yeongtong-Gu, Suwon 443-380, Korea; e-mail: radhej@naver.com

Indicates open access to non-subscribers at www.ajnr.org

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A6922

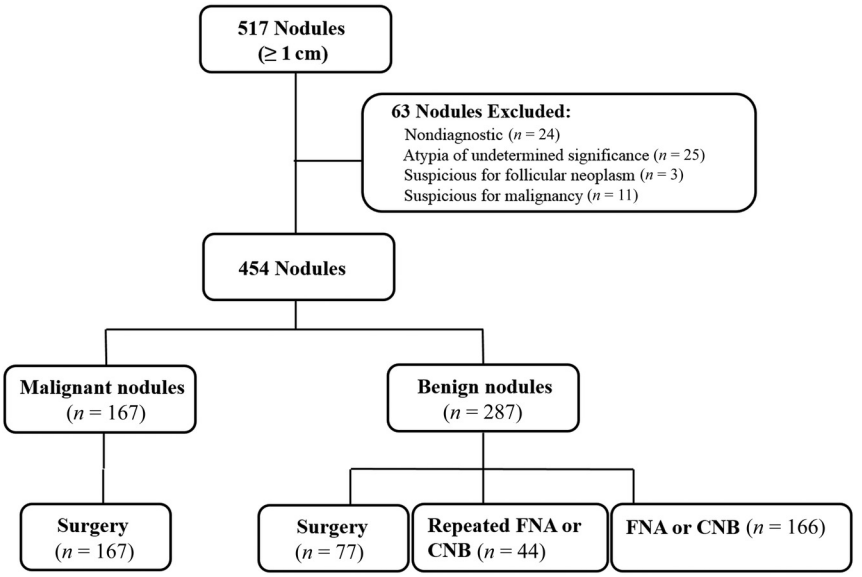

FIG 1. Flowchart shows the study participants. CNB indicates core-needle biopsy.

used for management of thyroid nodules since 2009;[13-15] therefore, evaluations are also needed to assess whether CAD systems can identify TIRADS categories and the risk of malignancy for each category. Two types of TIRADS have been used to manage thyroid nodules: pattern-based and point-based systems. Of these, the S-Detect CAD system takes a pattern-based TIRADS approach including the Korean Society of Thyroid Radiology-TIRADS (K-TIRADS) and the American Thyroid Association-TIRADS (ATA-TIRADS). A point-based TIRADS approach including the American College of Radiology–TIRADS is not yet available.

Therefore, we evaluated the diagnostic performance of the CAD US system in terms of detecting thyroid cancer based on pattern-based TIRADS (the K-TIRADS and ATA-TIRADS) and dichotomous outcome classification methods and compared its performance with that of an experienced radiologist.

MATERIALS AND METHODS

Patients

After obtaining institutional review board approval, written informed consent was obtained before US examinations from all patients. Between October 2018 and April 2019, four hundred fifty-three consecutive patients with 517 thyroid nodules (≥10 mm in diameter) who were referred to the department of radiology of our tertiary hospital for US-guided fine-needle aspiration (FNA) or US examination before a scheduled operation were initially enrolled. US-guided FNA was usually performed on a thyroid nodule exhibiting suspicious US features or on the largest nodule if no suspicious US feature was detected.[14] Data for 63 nodules were excluded because no final diagnoses were obtained (nondiagnostic, atypia of undetermined significance, and suspicions for follicular neoplasm and malignancy raised by FNA cytology but without surgical confirmation). Therefore, 372 patients with 454 thyroid nodules were finally included (83 males and 289 females; mean age, 49.5 years; range, 8–81 years; Fig 1).

Final diagnoses were determined from the cytopathologic results based on the Bethesda system and/or an operation. All malignant cases underwent thyroidectomy and were finally diagnosed by evaluation of surgical specimens. Benign nodules were diagnosed surgically or via benign core needle biopsy histology or cytologically benign FNA.

US Image Acquisition and Analyses

All thyroid US examinations were performed using a 3- to 12-MHz linear probe and a real-time US system (RS85A; Samsung Medison). Two experienced radiologists (E.J.H. and M.H.) with 14 and 10 years of clinical experience, respectively, performed all US examinations and US-guided biopsies.

The S-Detect 2 CAD system integrated into a commercially available US system was used to collect CAD data by the same radiologists. On the transverse image plane, an ROI was manually drawn around the target nodule.[10,11,16] The CAD system automatically outlined the contours of the mass and assessed the US features: composition (solid, partially cystic, or cystic); echogenicity (hyperechoic/isoechoic or hypoechoic); orientation (parallel or nonparallel); margins (well-defined, ill-defined, or microlobulated/spiculated); spongiform status; shape (ovoid to round or irregular); and calcifications (none, microcalcification, macrocalcification, or rim calcification). Finally, the CAD system provided a possible diagnosis using the TIRADS classification (based on the K-TIRADS and the ATA-TIRADS) or a dichotomous outcome classification (possibly benign/possibly malignant) (Fig 2).[13,14]

Gray-scale US images were retrospectively evaluated by the radiologist (E.J.H.) in terms of size, internal content, echogenicity, shape, orientation, margin, and the presence of calcification after at least 6 months, and the radiologist was blinded to all other data including the final histologic diagnoses.[14] The size, internal content, echogenicity, shape, orientation, margin, and calcifications were classified as described in previous reports.[13,14] On the basis of the US images, the nodules were classified according to the categories defined by the K-TIRADS and ATA-TIRADS, and a possible diagnosis was made by the radiologist.[13,14]

Statistical Analyses

Patient demographics, gray-scale US features, and dichotomous outcomes by the CAD system and radiologist were compared using the χ^2 or Fisher exact test. The Student *t* test was used to compare quantitative variables. The frequency and risk of malignancy according to each category of TIRADS were calculated as percentages. The associations between the categories of TIRADS and the final diagnoses were evaluated using the linear-by-linear association test.

The diagnostic abilities of the CAD system and the radiologist were assessed by calculating the sensitivities, specificities, positive

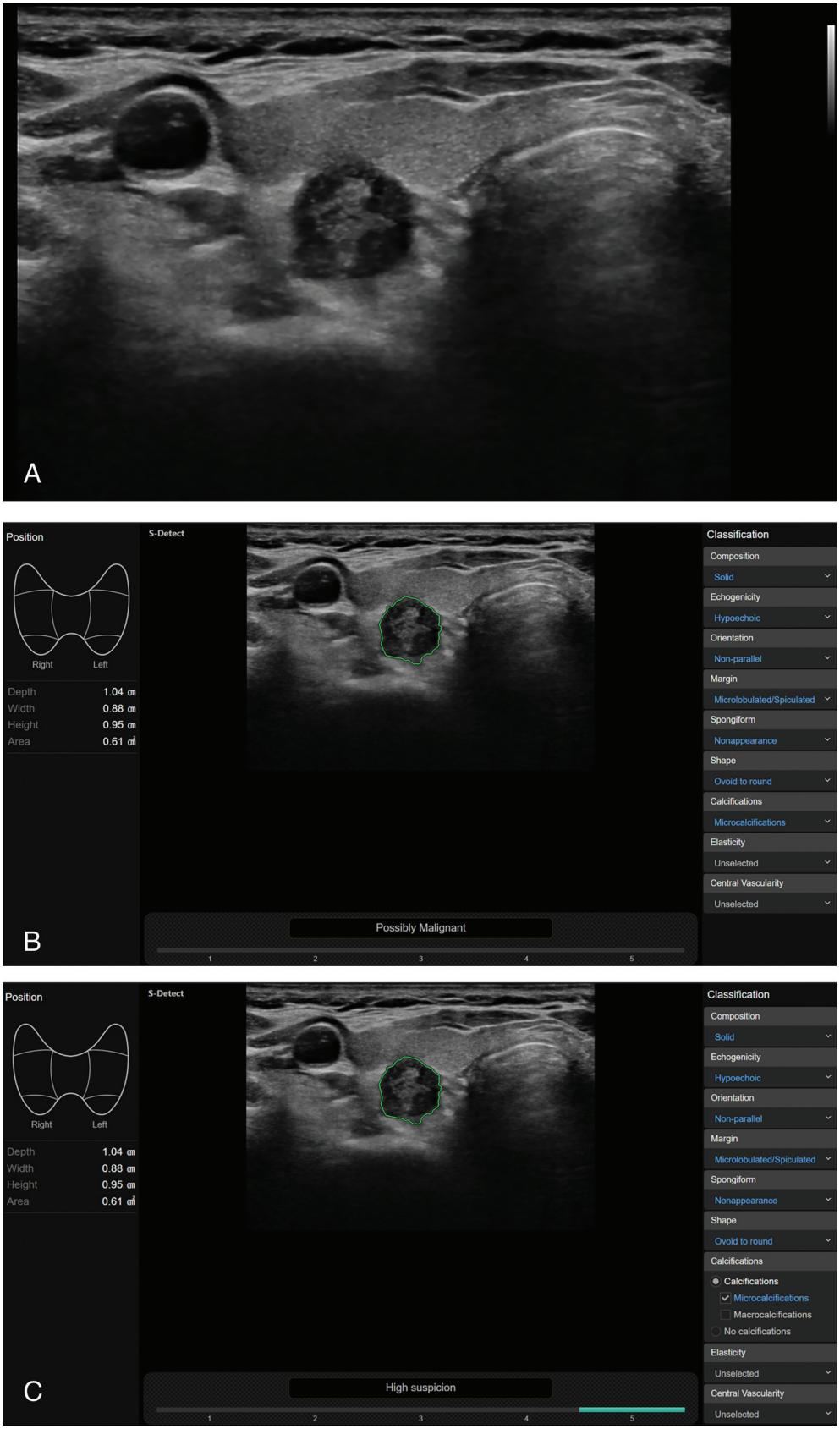

FIG 2. A US image of a thyroid nodule acquired with the S-Detect 2 CAD system. *A*, A solid hypoechoic nodule with suspicious US features is evident in the right thyroid gland. *B* and *C*, The CAD software automatically calculates the mass contours (*green contour*) and presents the US features on the right of the screen and a diagnosis based on the dichotomous outcome and TIRADS classification on the bottom.

Table 1: Sonographic features of thyroid nodules included in this study[a]

Characteristic	Benign Nodules ($n = 287$)	Malignant Nodules ($n = 167$)	Total ($n = 454$)	*P* Value
Diameter (mm)				.011
Mean [SD]	18.6 [SD, 10.7]	16.4 [SD, 7.5]	17.8 [SD, 9.7]	
Range	10.0–73.0	10.0–48.0	10.0–73.0	
Internal content				<.001
Solid	166 (57.8)	151 (90.4)	317 (69.8)	
Partially cystic	121 (42.2)	16 (9.6)	137 (30.2)	
Cystic	0	0	0	
Echogenicity				<.001
Hypoechogenicity	67 (23.3)	142 (85.0)	209 (46.0)	
Iso-/hyperechogenicity	220 (76.7)	25 (15.0)	245 (54.0)	
Shape				.143
Round-to-oval	278 (96.9)	157 (94.0)	435 (95.8)	
Irregular	9 (3.1)	10 (6.0)	19 (4.2)	
Orientation				<.001
Parallel	267 (93.0)	80 (47.9)	347 (76.4)	
Nonparallel	20 (7.0)	87 (52.1)	107 (23.6)	
Margin				<.001
Smooth	220 (76.7)	56 (33.5)	276 (60.8)	
Spiculated/microlobulated	5 (1.7)	85 (50.9)	90 (19.8)	
Ill-defined	62 (21.6)	26 (15.6)	88 (19.4)	
Calcification				<.001
None	256 (89.2)	49 (29.3)	305 (67.2)	
Microcalcification	8 (2.8)	102 (61.1)	110 (24.2)	
Macrocalcification	23 (8.0)	16 (9.6)	39 (8.6)	
Spongiform				0.534
Absence	285 (99.3)	167 (100.0)	452 (99.6)	
Presence	2 (0.7)	0	2 (0.4)	
CAD diagnosis				<.001
Possibly benign	235 (81.9)	31 (18.6)	266 (58.6)	
Possibly malignant	52 (18.1)	136 (81.4)	188 (41.4)	
Radiologist diagnosis				<.001
Possibly benign	275 (95.8)	30 (18.0)	305 (67.2)	
Possibly malignant	12 (4.2)	137 (82.0)	149 (32.8)	

[a] The numbers in parentheses are percentages.

Table 2: Risk of malignancy in each category based on different TIRADS using CAD and radiologist-based diagnosis[a]

TIRADS Classification	Category 1	Category 2	Category 3	Category 4	Category 5	*P* Value
CAD-based on K-TIRADS	0 (0/0)	4.8 (1/21)	12.1 (17/141)	20.9 (23/110)	69.2 (126/182)	<.001
Radiologist on K-TIRADS	0 (0/0)	0 (0/2)	6.0 (14/232)	37.5 (33/88)	90.9 (120/132)	<.001
CAD-based on ATA-TIRADS	7.1 (1/14)	11.4 (9/79)	10.6 (7/66)	25.5 (12/47)	65.7 (134/204)	<.001
Radiologist on ATA-TIRADS	0 (0/0)	7.3 (9/124)	5.5 (6/110)	26.3 (15/57)	90.6 (125/138)	<.001

[a] The numbers are percentages unless otherwise specified ; 9.6% (44 of 454) of nodules did not meet the criteria for any pattern using the ATA guidelines (isoechoic nodules with suspicious US features) and were classified as "not specified" by the CAD system, while the malignancy risk was calculated to be 9.1% (4 of 44).

predictive values, negative predictive values, and accuracy rates, and were compared using the McNemar test. Thyroid nodules requiring FNA as indicated by both sets of TIRADS recommendations were considered to indicate thyroid cancer (Online Supplemental Data).[13,14] We performed subgroup analyses of nodules at high and intermediate suspicion of cancer (as indicated by the FNA criteria). The unnecessary FNA rate was defined as the number of benign nodules among the FNA-required nodules (454 in total). Interobserver agreement between the CAD system and the radiologist in terms of the TIRADS and dichotomous outcome classifications was estimated using the κ coefficient. The κ level was defined as follows: <0.20, poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and >0.80, good agreement.

All statistical analyses were performed using SPSS for Windows (Version 25.0; IBM). The significance level was set at .05.

RESULTS

Clinical and Sonographic Features of Benign and Malignant Thyroid Nodules

The mean nodule diameter was 17.8 [SD, 9.7] mm (range, 10.0–73.0 mm). Of the 454 nodules, 287 (63.2%) were benign and 167 (36.8%) were malignant. Malignant nodules included 149 classic papillary thyroid carcinomas, 12 follicular-variant papillary thyroid carcinomas, 4 follicular carcinomas, and 2 medullary carcinomas.

Table 1 lists the US features of included nodules. The mean diameter of benign nodules was 18.6 [SD, 10.7] mm, which was statistically larger than that of malignant nodules (16.4 [SD, 7.5] mm; $P = .011$). Solid component, hypoechogenicity, nonparallel orientation, spiculated/microlobulated margins, and microcalcification were all significantly associated with thyroid cancer (all, $P < .001$). Diagnoses of "possibly malignant" by the CAD system and radiologist were significant in terms of detecting thyroid cancers (both, $P < .001$).

Table 3: Comparison of unnecessary FNA rates for thyroid cancer diagnosis in all nodules[a]

	No. of FNAs	No. of Malignant Nodules among FNA Nodules	No. of Benign Nodules among FNA Nodules	Unnecessary FNA Rate
CAD based on K-TIRADS (categories 3, 4, and 5)	85.5 (388/454)	42.0 (163/388)	58.0 (225/388)	49.6 (225/454)
CAD based on K-TIRADS (categories 4 and 5)	64.3 (292/454)	51.0 (149/292)	49.0 (143/292)	31.5 (143/454)
CAD based on ATA-TIRADS (categories 3, 4, and 5)	79.3 (360/454)	43.9 (158/360)	56.1 (202/360)	44.5 (202/454)
CAD based on ATA-TIRADS (categories 4 and 5)	55.3 (251/454)	57.5 (146/254)	42.5 (108/254)	23.8 (108/454)
CAD based on possible diagnosis	41.4 (188/454)	72.3 (136/188)	27.7 (52/188)	11.5 (52/454)
Radiologist K-TIRADS (categories 3, 4, and 5)	76.2 (346/454)	47.1 (163/346)	52.9 (183/346)	40.3 (183/454)
Radiologist K-TIRADS (categories 4 and 5)	48.5 (220/454)	69.5 (153/220)	30.5 (67/220)	14.8 (67/454)
Radiologist ATA-TIRADS (categories 3, 4, and 5)	68.3 (310/454)	48.4 (150/310)	51.6 (160/310)	35.2 (160/454)
Radiologist ATA-TIRADS (categories 4 and 5)	43.0 (195/454)	71.8 (140/195)	28.2 (55/195)	12.1 (55/454)
Radiologist based on possible diagnosis	32.8 (149/454)	91.9 (137/149)	8.1 (12/149)	2.6 (12/454)

[a] The numbers are percentages. Categories 5, 4, and 3 refer to high-, intermediate-, and low-suspicion nodules in each guideline.

Malignancy Risk according to CAD System and Radiologist and TIRADS Category Classifications

Table 2 lists the malignancy risk for each TIRADS category, classified by the CAD system and the radiologist. The malignancy risk for each K-TIRADS and ATA-TIRADS category determined by the radiologist matched the suggested malignancy risk, with the exception of a slightly higher risk of malignancy for the "very low suspicion" category of the ATA-TIRADS (7.3% [9 of 124] versus <3%). With the CAD system, the predicted probability of malignancy increased with the risk category ($P < .001$). However, when the CAD diagnosis was based on the ATA-TIRADS, the risk of malignancy did not match the suggested risk: It was higher for nodules that were benign and at very low, low, and intermediate suspicion but lower for nodules in the high-suspicion category. Overall, 9.6% (44 of 454) of nodules did not meet the criteria for any pattern using the ATA guidelines (isoechoic nodules with suspicious US features) and were classified as "not specified" by the CAD system. The malignancy risk was 9.1% (4 of 44).

Diagnostic Performance of the CAD System and Radiologist Based on TIRADS Classifications and Dichotomous Outcomes

The Online Supplemental Data summarize thyroid cancer diagnostic performance by the CAD system and the radiologist based on the TIRADS and dichotomous outcome classifications. The sensitivity and negative predictive values were highest for radiologist K-TIRADS and CAD K-TIRADS, followed by CAD ATA-TIRADS, radiologist ATA-TIRADS, radiologist's diagnosis, and CAD diagnosis. The specificity and positive predictive values were highest for the radiologist's diagnosis, followed by the CAD diagnosis, radiologist ATA-TIRADS, radiologist K-TIRADS, CAD ATA-TIRADS, and CAD K-TIRADS.

The TIRADS classifications had significantly higher diagnostic sensitivities but lower specificities compared with dichotomous outcome classifications, while the latter had higher specificities (all, $P < .001$). The diagnostic sensitivities of the CAD systems using the TIRADS classification and dichotomous outcomes did not differ between the CAD systems and radiologist (97.6% versus 97.6%, $P > .999$, for the K-TIRADS; 94.6% versus 89.8%, $P = .077$, for the ATA-TIRADS; and 81.4% versus 82.0%, $P > .999$, for the possible diagnosis, respectively), while the specificity and accuracy were significantly lower for the CAD systems compared with the radiologist (21.6% versus 36.2%; 29.6% versus 44.3%; and 81.9% versus 95.8%, all $P < .001$, respectively, and 49.6% versus 58.8%; 53.5% versus 61.0%; and 81.7% versus 90.7%, respectively; all, $P < .001$).

When we used the FNA criterion to evaluate nodules at high and intermediate suspicion of malignancy, the diagnostic specificity and accuracy of the CAD system increased; however, the diagnostic performance of the TIRADS classifications (compared with the dichotomous outcome classification) was similar to that of the overall diagnostic performance.

Comparison of Unnecessary FNA Rates

The unnecessary FNA rate was the lowest for the radiologist's diagnosis, followed by the CAD diagnosis, radiologist ATA-TIRADS, radiologist K-TIRADS, CAD ATA-TIRADS, and CAD K-TIRADS (Table 3). The dichotomous outcome classification yielded a lower unnecessary FNA rate than the TIRADS classification, particularly by the radiologist.

Interobserver Agreement between the CAD System and the Radiologist

The dichotomous outcome agreement for the CAD system and the radiologist was 83.0% (377/454). The extent of interobserver agreement was substantial ($\kappa = 0.640$) for the dichotomous outcomes and fair-to-moderate to the K-TIRADS and the ATA-TIRADS classifications ($\kappa = 0.356$ and 0.402, respectively, Table 4).

DISCUSSION

Our results revealed that the diagnostic performance of the CAD system varies with the TIRADS and dichotomous outcome classifications. Dichotomous outcomes revealed significantly higher specificity, positive predictive values, and accuracy for detecting

Table 4: Interobserver variability of US characteristics between radiologist and CAD systems[a]

	κ Value
K-TIRADS (CAD vs radiologist)	0.356
ATA-TIRADS (CAD vs radiologist)	0.402
Possible diagnosis (CAD vs radiologist)	0.640

[a] The extent of interobserver agreement between the CAD system and the radiologist was calculated using the Cohen κ value.

thyroid cancer, an outcome associated with a reduction in the unnecessary FNA rates. However, the TIRADS classification achieved higher sensitivity and negative predictive values, which increased unnecessary FNA rates. Clinicians should be aware of these particular strengths and weaknesses of the CAD system in the management of thyroid nodules.

The use of high-resolution US, combined with increased medical surveillance and access to health care services, has markedly increased the detection of thyroid nodules and the number of FNAs.[15,16] Therefore, radiologists who frequently interpret thyroid US images are concerned about how to report nodules and on which nodules to perform FNA. Since 2009, the use of the TIRADS classification system has been recommended to improve consistency across practices and institutions and to decrease unnecessary FNAs.[13-15] Several professional groups, including the American Thyroid Association and the Korean Society of Thyroid Radiology, have proposed the ATA-TIRADS and K-TIRADS, respectively, and have recommended FNA criteria in conjunction with the nodule size and TIRADS category.[13,14] In keeping with this international trend, the currently available CAD system provides both TIRADS classifications and dichotomous outcomes. This CAD system is based on training of a deep learning algorithm using 4916 nodules from 3 different institutions.[12] We found that the risk of malignancy significantly increased with the higher risk categories when the TIRADS category was assigned by the CAD system; however, the calculated prevalence and risk in each category differed depending on whether the CAD-based or radiologist-based method was used. The CAD system overestimated the number of TIRADS category 5 (highly suspicious) nodules and underestimated the risk of malignancy in TIRADS category 5 compared with the radiologist. Therefore, CAD users should be aware that the risk of malignancy differs by category between the CAD- and radiologist-based methods.

In terms of system diagnostic performance, similar sensitivity scores have been reported for CAD- and radiologist-based assessments.[9-11,16] However, reduced specificity and accuracy have been reported for the CAD-based system.[9-11,16] In agreement with these findings, we observed lower specificity and accuracy for the CAD system compared with the radiologist (81.9% versus 95.8% and 81.7% versus 90.7%, respectively) and similar sensitivity (81.4% versus 82.0%) for the detection of thyroid cancer. Furthermore, we present the first assessment of the diagnostic ability of the TIRADS classification of the CAD system. We found that the TIRADS classification had significantly higher diagnostic sensitivities (94.6%–97.6% versus 81.4%) but lower specificities (21.6%–29.6% versus 81.9%) compared with dichotomous outcomes, which increase unnecessary FNA rates (44.5%–49.6% versus 11.5%). The false-positive rate was higher for the CAD system, while the false-negative rate was not significantly changed. However, these differences were reduced when the FNA criteria for nodules at high and intermediate suspicion were applied. Our study identified only fair-to-moderate agreement between the CAD system and the radiologist's TIRADS classifications, which highlights a limitation of the current CAD system. The interobserver agreement between the CAD system and the radiologist in terms of the margins and calcifications was the lowest but remained fair-to-moderate (κ = 0.390 and 0.448, respectively), reducing the overall system accuracy. A recent blinded multicenter study similarly reported that the inter- and intraobserver agreement (using a US classification system) were 0.34–0.44 and 0.33–0.54, respectively, among thyroid imaging experts.[16] Therefore, CAD users should be aware of the strengths and weaknesses associated with thyroid cancer diagnosis using commercially available CAD systems.

Our study revealed important design issues for an AI-based thyroid cancer CAD system. Previous studies have relied on a simple classification model (benign/malignant) without the inclusion of US features.[4–7] However, several US features are strongly associated with thyroid cancer, and a simple classification system cannot incorporate the influence of these US features on the final diagnosis in convolutional neural network deep learning models.[17,18] Therefore, the currently available CAD system was designed to report information about US features in addition to the possible diagnosis to help inform convolutional neural network deep learning models and infer a conclusion. Such a system could offer great advantages. Less experienced operators find it difficult to accurately recognize and consistently interpret US features, so an AI-based CAD system would improve standardization and ultimately reduce unnecessary FNAs.[11] However, on the contrary, the dichotomous AI prediction showed relatively high specificity and positive predictive values that, in fact, match or exceed nearly all permutations of testing performed in this study with the exception of expert radiologist-based diagnosis. This finding implies that even if a radiologist were able to perfectly score a lesion based on the TIRADS classification, the dichotomous AI prediction may help reduce false-positive FNAs compared with TIRADS-based triage. Further improvements and validations are required on this issue.

Our study had certain limitations. First, we included nodules that had been referred to US-guided FNA or US examination before a scheduled operation. Therefore, the proportion of malignancies was high, and the diagnostic performance of the system might differ in a general population. Second, the radiologist's diagnoses were based on personal experience, so a less experienced radiologist might have reported differently. This feature may influence the generalizability of this study. Third, the CAD data were obtained by the same radiologist who performed US. However, the CAD data were semiautomatically obtained and the radiologist retrospectively assessed the US findings after at least 6 months while blinded to other data, so this process minimized bias. Fourth, the clinical impact of the CAD system might differ slightly in real-world practice. Further research using a prospective study design is required in a general population.

CONCLUSIONS

The diagnostic performance of the CAD system differs depending on the TIRADS and dichotomous outcome classifications and

compared with radiologist-based assessments. Clinicians should be aware of the strengths and weaknesses of the CAD system.

Disclosures: Eun Ju Ha—*RELATED*: *Grant*: Research Fund for Young Scientists.

REFERENCES

1. Park SH, Kressel HY. **Connecting technological innovation in artificial intelligence to real-world medical practice through rigorous clinical validation: what peer-reviewed medical journals could do.** *J Korean Med Sci* 2018;33:e152 CrossRef Medline
2. Park SH, Han K. **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.** *Radiology* 2018;286:800–09 CrossRef Medline
3. Park SH. **Regulatory approval versus clinical validation of artificial intelligence diagnostic tools.** *Radiology* 2018;288:910–11 CrossRef Medline
4. Li X, Zhang S, Zhang Q, et al. **Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study.** *Lancet Oncol* 2019;20:193–201 CrossRef Medline
5. Chi J, Walia E, Babyn P, et al. **Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network.** *J Digit Imaging* 2017;30:477–86 CrossRef Medline
6. Buda M, Wildman-Tobriner B, Castor K, et al. **Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images.** *Ultrasound Med Biol* 2020;46:415–21 CrossRef Medline
7. 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 CrossRef Medline
8. Kim DW, Jang HY, Kim KW, et al. **Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers.** *Korean J Radiol* 2019;20:405–10 CrossRef Medline
9. Choi YJ, Baek JH, Park HS, et al. **A computer-aided diagnostic system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment.** *Thyroid* 2017;27:546–52 CrossRef Medline
10. Yoo YJ, Ha EJ, Cho YJ, et al. **A computer-aided diagnostic system for thyroid nodules on ultrasonography: initial clinical experience.** *Korean J Radiol* 2018;28:1532–37 CrossRef Medline
11. Jeong EY, Kim HL, Ha EJ, et al. **Computer-aided diagnostic system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators.** *Eur Radiol* 2019;29:1978–85 CrossRef Medline
12. Park VY, Han K, Seong YK, et al. **Diagnosis of thyroid nodules: performance of a deep learning convolutional neural network model vs radiologists.** *Sci Rep* 2019;9:17843 CrossRef Medline
13. Haugen BR. **2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: what is new and what has changed?** *Cancer* 2017;123:372–81 CrossRef Medline
14. Shin JH, Baek JH, Chung J, et al; Korean Society of Thyroid Radiology (KSThR) and Korean Society of Radiology. **Ultrasonography Diagnosis and Imaging-Based Management of Thyroid Nodules: Revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations.** *Korean J Radiol* 2016;17:370–95 CrossRef Medline
15. Ha EJ, Baek JH, Na DG. **Risk stratification of thyroid nodules on ultrasonography: current status and perspectives.** *Thyroid* 2017;27:1463–68 CrossRef Medline
16. Kim HL, Ha EJ, Han M. **Real-world performance of computer-aided diagnostic system for thyroid nodules using ultrasonography.** *Ultrasound Med Biol* 2019;45:2672–78 CrossRef Medline
17. Persichetti A, Stasio E, Coccaro C, et al. **Inter- and intraobserver agreement in the assessment of thyroid nodule ultrasound features and classification systems: a blinded multicenter study.** *Thyroid* 2020;30:237–42 CrossRef Medline
18. Ha EJ, Baek JH, Na DG. **Deep convolutional neural network models for the diagnosis of thyroid cancer.** *Lancet Oncol* 2019;20:e130 CrossRef Medline