European Radiology (2019) 29:1978–1985
https://doi.org/10.1007/s00330-018-5772-9

HEAD AND NECK

Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators

Eun Young Jeong[1] · Hye Lin Kim[1] · Eun Ju Ha[1] · Seon Young Park[1] · Yoon Joo Cho[1] · Miran Han[1]

Received: 27 July 2018 / Revised: 24 August 2018 / Accepted: 18 September 2018 / Published online: 22 October 2018

Abstract

Purpose To evaluate the diagnostic performance and reproducibility of a computer-aided diagnosis (CAD) system for thyroid cancer diagnosis using ultrasonography (US) based on the operator's experience.

Materials and methods Between July 2016 and October 2016, 76 consecutive patients with 100 thyroid nodules (≥ 1.0 cm) were prospectively included. An experienced radiologist performed the US examinations with a real-time CAD system integrated into the US machine, and three operators with different levels of US experience (0–5 years) independently applied the CAD system. We compared the diagnostic performance of the CAD system based on the operators' experience and calculated the interobserver agreement for cancer diagnosis and in terms of each US descriptor.

Results The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of the CAD system were 88.6, 83.9, 81.3, 90.4, and 86.0%, respectively. The sensitivity and accuracy of the CAD system were not significantly different from those of the radiologist ($p > 0.05$), while the specificity was higher for the experienced radiologist ($p = 0.016$). For the less-experienced operators, the sensitivity was 68.8–73.8%, specificity 74.1–88.5%, PPV 68.9–73.3%, NPV 72.7–80.0%, and accuracy 71.0–75.0%. The less-experienced operators showed lower sensitivity and accuracy than those for the experienced radiologist. The interobserver agreement was substantial for the final diagnosis and each US descriptor, and moderate for the margin and composition.

Conclusions The CAD system may have a potential role in the thyroid cancer diagnosis. However, operator dependency still remains and needs improvement.

Key Points

- *The sensitivity and accuracy of the CAD system did not differ significantly from those of the experienced radiologist (88.6% vs. 84.1%, p = 0.687; 86.0% vs. 91.0%, p = 0.267) while the specificity was significantly higher for the experienced radiologist (83.9% vs. 96.4%, p = 0.016).*
- *However, the diagnostic performance varied according to the operator's experience (sensitivity 70.5–88.6%, accuracy 72.0–86.0%) and they were lower for the less-experienced operators than for the experienced radiologist.*
- *The interobserver agreement was substantial for the final diagnosis and each US descriptor and moderate for the margin and composition.*

Keywords Artificial intelligence · Fine-needle aspiration · Thyroid nodule · Thyroid cancer · Ultrasonography

Eun Young Jeong and Hye Lin Kim contributed equally to this work.

✉ Eun Ju Ha
radhej@naver.com

[1] Department of Radiology, Ajou University School of Medicine, Wonchon-Dong, Yeongtong-Gu, Suwon 443-380, South Korea

Abbreviations

AUC	Area under receiver operating characteristic curve
CAD	Computer-aided diagnosis
CI	Confidence interval
FNA	Fine-needle aspiration
NPV	Negative predictive value
PPV	Positive predictive value
PTC	Papillary thyroid carcinoma

ROC Receiver operating characteristic
US Ultrasonography

Introduction

Ultrasonography (US) is the most accurate imaging modality for the differential diagnosis of thyroid nodules. Therefore, current practice guidelines recommend that US should be performed in all patients with a suspected thyroid nodule on physical examination, nodular goiter, or radiographic abnormality on another imaging study suggesting a thyroid nodule [1, 2]. However, since US image interpretation is operator-dependent and interobserver variability is moderate to substantial [3–7], the diagnostic performance of US ranges from 40.3 to 100.0% sensitivity and 50.0 to 100.0% specificity [3–5, 7]. Therefore, unnecessary fine-needle aspiration (FNA) and even diagnostic surgery are common, placing a significant burden on healthcare systems and creating patient anxiety [3–7].

A computer-aided diagnosis (CAD) system was recently introduced for accurate, consistent interpretation of the US features of thyroid nodules [8–12]. Several studies have found that the CAD system affords a diagnostic accuracy similar to that of an experienced radiologist and it could offer support for decision-making in thyroid cancer diagnosis [9, 10]. However, since these studies were performed by experienced radiologists, there are problems regarding the actual interpretation of the data and the reproducibility of the CAD results, especially when performed by less-experienced operators. The current US CAD system involves four main steps: targeting the area of a lesion manually; selecting one of candidates that the CAD presents; editing the boundary of the chosen candidate (this step can be omitted when the candidate has a fine boundary); and letting the CAD system analyze the US image and output its features [10]. As these semi-automated steps require experience of the operators, the actual diagnostic performance and reproducibility of the CAD system based on operator experience should be evaluated to optimize its utility.

Therefore, this study evaluated the diagnostic performance and reproducibility of the CAD system regarding a US thyroid cancer diagnosis based on the operator's experience.

Materials and methods

Patients

This study was approved by our institutional review board and written informed consent was obtained from all patients before they underwent US. Patient data were prospectively collected from a single tertiary hospital. Patients were usually referred from the primary medical centers due to a suspicious thyroid nodule or from the other department due to a thyroid nodule detected on another imaging study.

Between July 2016 and October 2016, a total of 85 consecutive patients with 109 thyroid nodules (≥ 1.0 cm) who underwent US-guided biopsy or US examination before the scheduled surgery were initially enrolled in this study. US-guided FNA or core needle biopsy (CNB) was usually performed on a thyroid nodule with suspicious US features or on the largest nodule if no suspicious US feature was detected based on the guidelines [2]. Among them, 9 nodules were excluded because a final diagnosis was not obtained (nondiagnostic ($n = 3$), atypia of undetermined significance ($n = 4$), and suspicion for malignancy ($n = 2$)). A total of 76 consecutive patients with 100 thyroid nodules were finally included in this study (23 males, 53 females; mean age 46.0 years (range 14–75 years)) (Fig. 1).

Final diagnoses were determined from the cytopathological results based on the Bethesda system and/or surgery. A malignant nodule was diagnosed when malignancy was confirmed in the surgical specimen. A benign nodule was diagnosed when either (1) benign status was confirmed in the surgical specimen or (2) it was cytologically benign on FNA or core needle biopsy histology. Cytological results from nondiagnostic lesions and lesions of indeterminate significance (atypia of undetermined significance and suspicions of follicular neoplasm and malignancy) without surgical confirmation were excluded.

US image acquisition and analysis

All US examinations were performed using a 5–12-MHz linear probe and a real-time CAD US system (RS80A; Samsung Medison). The real-time CAD system (S-Detect for Thyroid; Samsung Medison) was integrated into the US system. A radiologist (E.J.H.) specializing in thyroid imaging with 10 years of clinical experience in the performance of thyroid US (and evaluation of its data) performed all US examinations.

Grayscale US images were evaluated by the radiologist (E.J.H) with reference to the Korean guidelines for size, internal content, echogenicity, shape, orientation, margin, and the presence or absence of calcifications [2]. The nodule contents were categorized by reference to the cystic proportion: solid (no obvious cystic content), predominantly solid (< 50% cystic), predominantly cystic (> 50% cystic), or cystic (pure cyst or nearly entirely cystic content). The echogenicity was categorized as marked hypoechogenicity (when a nodule showed a relatively hypoechoic pattern with regard to the adjacent strap muscle), mild hypoechogenicity (when a nodule showed a relatively hypoechoic pattern with regard to the normal thyroid parenchyma), isoechogenicity (when a nodule showed an isoechoic pattern with regard to the normal thyroid parenchyma), or hyperechogenicity (when a nodule showed a relatively

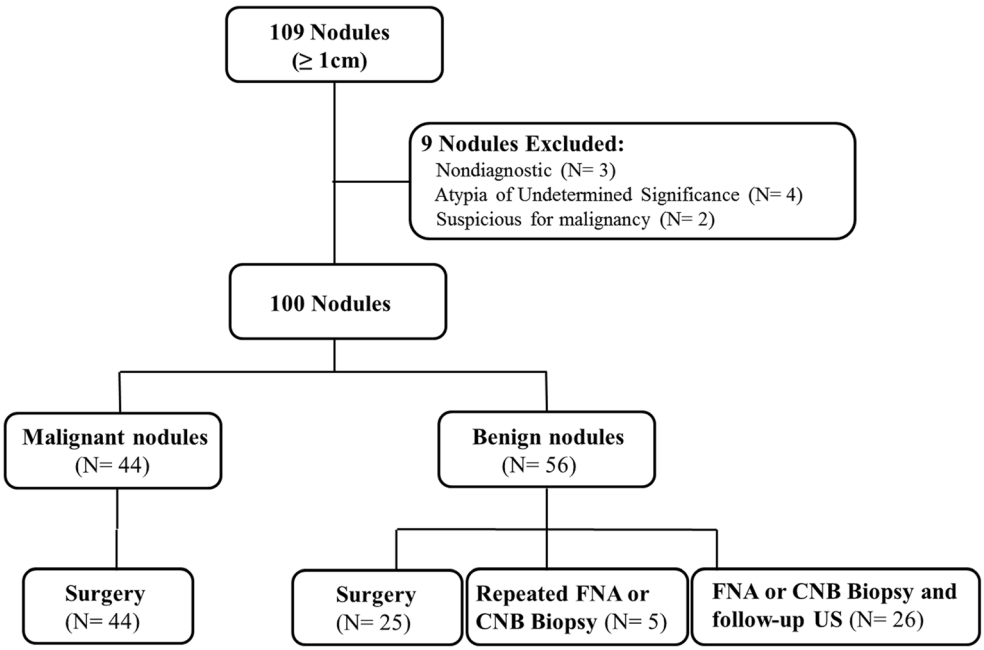

Fig. 1 Flowchart of the study

echogenic pattern with regard to the normal thyroid parenchyma). Shape was categorized as ovoid-to-round or irregular, and orientation was categorized as parallel (when the anteroposterior diameter of the nodule was equal to or less than the transverse or longitudinal diameter) or non-parallel (when the anteroposterior diameter of the nodule exceeded the transverse and longitudinal diameter in the transverse and longitudinal plane, respectively). The margins were categorized as smooth, spiculated/microlobulated, or ill-defined. The spiculated/microlobulated margin was defined when the margin of any portion of a nodule is obviously discernible, but non-smooth edge showing spiculation, microlobulation, or jagged appearance. The ill-defined margin was defined when the border of the nodule is poorly demarcated which cannot be obviously differentiated from adjacent thyroid tissue. A smooth margin was defined when the nodule shows an obviously discernible smooth edge. Calcification was classified as microcalcification (tiny, punctate echogenic foci ≤ 1 mm in diameter, with or without posterior shadowing), macrocalcification (echogenic foci > 1 mm in diameter), or rim calcification (peripheral curvilinear or eggshell-like calcification). Spongiform appearance of a nodule was defined as the aggregation of multiple microcystic components in more than 50% of the isoechoic partially cystic nodule.

CAD image acquisition and analysis

The CAD data were obtained by four operators with different levels of US experience (E.J.H., a staff radiologist with 10 years of experience in the performance and evaluation of thyroid US; S.Y.P., a fellow radiologist specializing in non-thyroid US with 5 years of experience in US; H.L.K., a 2nd year radiology resident who had 1 year of experience with US; and E.Y.J., a medical student who had no experience in US). They assessed the CAD data without any information on the other results. We only used still images for the evaluation and cine loops were not used in this study. An experienced radiologist recorded the still images of each nodule in real-time US and provided images to the operators. All the sectional images used to set the region of interest for was the same for all operators in each nodules.

Before starting the study, the basic methods were discussed to establish a consensus on the use of the CAD system. The CAD data were obtained from transverse planes by manually setting a region of interest around the lesion. The software calculated the mass contours automatically (thereby distinguishing the mass from normal thyroid tissue) and evaluated the US features of the mass, including its composition (solid, partially cystic, or cystic), shape (oval-to-round or irregular), orientation (parallel or non-parallel), margins (well-defined, ill-defined, or spiculated), echogenicity (hyperechoic/isoechoic or hypoechoic/markedly hypoechoic), and spongiform status. In terms of the margins, the operator chose one of the three to four options suggested by the software and edited the margin manually if needed. The CAD system ultimately diagnosed the nodules as benign or malignant using the abovementioned descriptors (Fig. 2).

Data and statistical analysis

Differences in sonographic features (size, internal content, echogenicity, shape, orientation, margin, and the presence or absence of calcifications) and CAD diagnosis (probably benign and probably malignant) between benign and

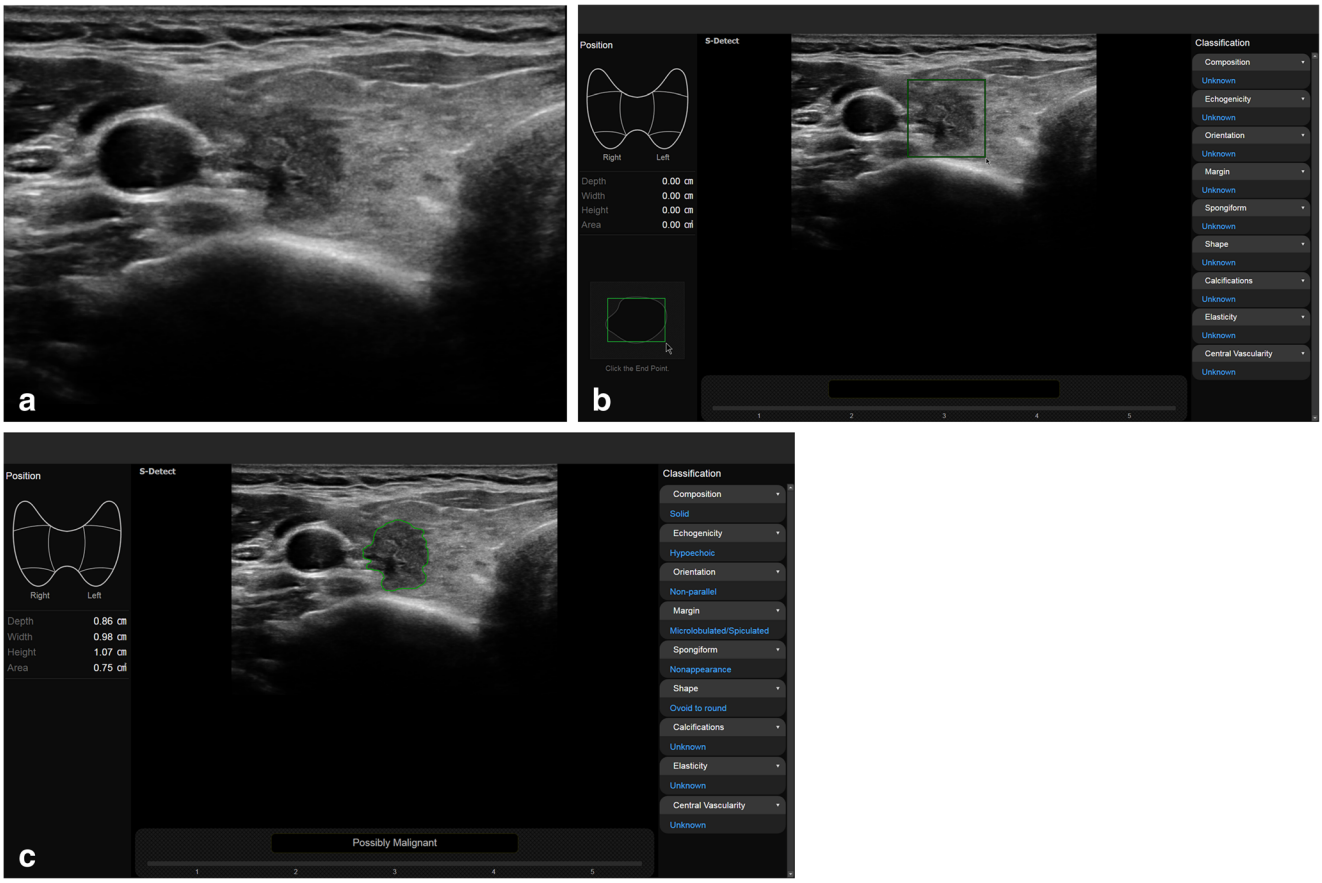

Fig. 2 An ultrasonography (US) image of a thyroid nodule acquired with the computer-aided diagnosis (CAD) system. **a** A solid hypoechoic nodule with suspicious US features is evident in the right thyroid gland. **b** A region of interest is manually drawn around the lesion. **c** The CAD software automatically calculates the mass contours and presents the US features on the right of the screen, and a possible diagnosis as a malignant nodule on the bottom

malignant thyroid nodules were evaluated using the χ^2 or Fisher's exact test. Student's *t* test was used to compare a quantitative variable (size).

The diagnostic performance of the CAD system based on the experience level of the operators was evaluated by calculating the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. The diagnostic performance of the CAD system was compared using the McNemar test. The area under the receiver operating characteristic (ROC) curve with the 95% confidence interval (CI) was calculated for the CAD system obtained with four different operators for thyroid cancer diagnosis. The extent of interobserver variability (the κ value) among the CAD final diagnoses obtained by different operators, and in terms of the descriptions of the US characteristics, was determined. The Fleiss multirater kappa statistic was used to obtain an overall κ value among multiple observers. The level of agreement was defined as follows $\kappa < 0.0$, poor; $0.0 < \kappa < 0.20$, slight; $0.21 < \kappa < 0.40$, fair; $0.41 < \kappa < 0.60$, moderate; $0.61 < \kappa < 0.80$, substantial and $0.81 < \kappa < 1.00$, almost perfect [13, 14].

All statistical analyses were performed using SPSS for Windows (ver. 23.0; IBM Corp.), MedCalc for Windows (ver. 15.0; MedCalc), and R 3.4.1 for Windows software (R Development Core Team). A significant difference was defined as $p < 0.05$.

Results

Demographic data and US features predicting malignant thyroid nodules

The mean nodule diameter was 1.7 ± 0.8 cm (range 1.0–4.6 cm). In the final diagnoses, 56 (56.0%) of the 100 nodules were benign and 44 (44.0%) were malignant. All of the final diagnoses of malignant nodules were confirmed after surgical resection, and included 43 papillary thyroid carcinomas (PTCs), which in turn included 41 classic PTCs, 1 follicular variant PTC, and 1 oncocytic variant PTC, and 1 follicular carcinoma. The 25 surgically confirmed benign nodules were all nodular hyperplasias.

Table 1 summarizes the US features of the benign and malignant nodules. The mean diameter of the benign nodules was 1.8 ± 0.8 cm, which exceeded that of the

Table 1 Sonographic features of the benign and malignant thyroid nodules

	Benign ($n = 56$)	Malignant ($n = 44$)	p value
Size (cm)			0.031
Mean ± SD	1.8 ± 0.8	1.5 ± 0.8	
Range	1.0–3.9	1.0–4.6	
Internal content			< 0.001
Solid	25 (44.6)	39 (88.6)	< 0.001
Predominantly solid	21 (37.5)	4 (9.1)	0.001
Predominantly cystic	9 (16.1)	1 (2.3)	0.040
Cystic	1 (1.8)	0 (0.0)	1.000
Echogenicity			< 0.001
Marked hypoechogenic	3 (5.4)	5 (11.4)	0.295
Hypoechogenic	15 (26.8)	32 (72.7)	< 0.001
Isoechogenic	38 (67.9)	7 (15.9)	< 0.001
Hyperechogenic	0 (0.0)	0 (0.0)	–
Shape			0.592
Round-to-oval	52 (92.9)	42 (95.5)	
Irregular	4 (7.1)	2 (4.5)	
Orientation			< 0.001
Parallel	52 (92.9)	27 (61.4)	
Non-parallel	4 (7.1)	17 (38.6)	
Margin			< 0.001
Smooth	49 (87.5)	11 (25.0)	<0.001
Spiculated/microlobulated	0 (0.0)	25 (56.8)	< 0.001
Ill-defined	7 (12.5)	8 (18.2)	0.430
Calcification			< 0.001
None	50 (89.3)	10 (22.7)	< 0.001
Microcalcification	3 (5.4)	30 (68.2)	< 0.001
Macrocalcification	3 (5.4)	4 (9.1)	0.696
Rim calcification	0 (0.0)	0 (0.0)	–
Spongiform appearance			–
Presence	0 (0.0)	0 (0.0)	
Absence	0 (0.0)	0 (0.0)	
Computer-aided diagnosis			< 0.001
Probably benign	47 (83.9)	5 (11.4)	
Probably malignant	9 (16.1)	39 (88.6)	

The numbers in parentheses are percentages. Student's t test was used to compare a quantitative variable (size). χ^2 or Fisher's exact test was used to compare the others

CAD computer-aided diagnosis

malignant nodules (1.5 ± 0.8 cm; $p = 0.031$). Alongside the US features, including a solid component, marked hypoechogenicity, non-parallel orientation, spiculated margins, and microcalcification, the "probably malignant" diagnostic option of the CAD system was a significant factor influencing the rate of detection of thyroid cancers ($p < 0.001$).

Diagnostic performance of the CAD system and the experienced radiologist

Table 2 shows the diagnostic performances of the experienced radiologist and the CAD system for thyroid cancer. The sensitivity of the CAD system did not differ significantly from that of the experienced radiologist (88.6% vs. 84.1%, $p = 0.687$); while the specificity of the CAD system was significantly lower than that of the experienced radiologist (83.9% vs. 96.4%, $p = 0.016$). Diagnostic accuracy did not differ significantly between the CAD system and the radiologist (86.0% vs. 91.0%, $p = 0.267$).

Diagnostic performance of the CAD system obtained by four different operators

Table 3 summarizes the diagnostic performance of the CAD system obtained by four different operators with different levels of experience at performing thyroid US. The sensitivity, specificity, PPV, NPV, and accuracy of the CAD system were 88.6, 83.9, 81.3, 90.4, and 86.0%, respectively, for the experienced radiologist, while for the less-experienced operators the sensitivity was 70.5–75.0%, specificity 73.2–80.4%, PPV 67.4–73.8%, NPV 75.0–78.8%, and accuracy 72.0–76.0%. The diagnostic sensitivity of the CAD system for the experienced radiologist were higher than those for the less-experienced operators ($p = 0.008$, 0.070, and 0.039, respectively). The diagnostic accuracy of the CAD system for the

Table 2 Diagnostic performances of the experienced radiologist and the computer-aided diagnosis system for thyroid cancer

Diagnostic measure	Experienced radiologist	Computer-aided diagnosis	*p* value
Sensitivity	84.1 (37/44)	88.6 (39/44)	0.687
Specificity	96.4 (54/56)	83.9 (47/56)	0.016
Positive predictive value	94.9 (37/39)	81.3 (39/48)	
Negative predictive value	88.5 (54/61)	90.4 (47/52)	
Accuracy	91.0 (91/100)	86.0 (86/100)	0.267

The *p* value is that of the experienced radiologist versus the CAD system. The McNemar test was used for the statistical analysis

experienced radiologist was significantly higher than those for the less-experienced operators (all $p < 0.05$). The specificity of the CAD system was higher for the experienced radiologist, although not significantly different ($p = 0.754$, 0.180, and 0.109, respectively).

Figure 3 shows the ROC curves for the CAD system obtained by the four different operators for thyroid cancer diagnosis. The AUC was 0.863 (95% CI, 0.780–0.923) for the experienced radiologist, which was significantly higher than that for the less-experienced radiologist (0.754 (0.648–0.835), 0.741 (0.644–0.824), and 0.718 (0.620–0.804), respectively ($p = 0.008$, 0.007, and 0.002, respectively)).

Interobserver agreement for the CAD system with four different operators

Table 4 presents the interobserver agreement for the CAD system for four different operators. The extent of interobserver agreement for the final diagnosis was substantial ($\kappa = 0.658$). In terms of each US descriptor, there was substantial agreement for spongiform ($\kappa = 0.797$), echogenicity ($\kappa = 0.723$), and orientation ($\kappa = 0.710$), and moderate agreement for margin ($\kappa = 0.479$) and composition ($\kappa = 0.444$). For the agreement regarding shape, the CAD data reported "ovoid-to-round" margins in all cases for all observers, so that the kappa value could not be calculated.

Discussion

This study demonstrated that the sensitivity and accuracy of the CAD system for thyroid cancer diagnosis were not significantly different from those of the experienced radiologist. However, operator dependency still remains since they varied according to the operator's experience and were lower for the less-experienced operators than for the experienced radiologist.

Ultrasound is the primary diagnostic tool used for assessing the malignancy risk of thyroid nodules and aiding decision-making regarding the use of FNA [1, 2]. Despite its importance, many studies have shown that the diagnostic performance of US is variable and affected by the operator's experience [3–7]. This indicates that the diagnosis using US is operator-dependent, where less-experienced operators may make inappropriate decisions regarding the use of FNA. Therefore, a new CAD system using artificial intelligence was expected to improve the diagnostic performance of US and decrease the interobserver variability via use of a semi-automated workflow [8–12, 15]. Regarding the diagnostic performance of the CAD system, several studies have reported comparable diagnostic performance between the CAD system and experienced radiologists [9, 10]. Choi et al reported a sensitivity of up to 88.4% and suggested that the CAD system was useful for ruling out thyroid malignancy on US and was an easy method for determining the need for an FNA biopsy

Table 3 Diagnostic performance of the computer-aided diagnosis system obtained with four different operators

Diagnostic measure	Operator 1	Operator 2	Operator 3	Experienced radiologist	*p* value [1]	*p* value [2]	*p* value [3]
Sensitivity	70.5 (31/44)	75.0 (33/44)	70.5 (31/44)	88.6 (39/44)	0.008	0.070	0.039
Specificity	80.4 (45/56)	73.2 (41/56)	73.2 (41/56)	83.9 (47/56)	0.754	0.180	0.109
Positive predictive value	73.8 (31/42)	68.8 (33/48)	67.4 (31/46)	81.3 (39/48)			
Negative predictive value	77.6 (45/58)	78.8 (41/52)	75.0 (41/54)	90.4 (47/52)			
Accuracy	76.0 (76/100)	74.0 (74/100)	72.0 (71/100)	86.0 (86/100)	0.031	0.017	0.004

CAD computer-aided diagnosis

[1] *p* value of the CAD system obtained by operator 1 versus the experienced radiologist

[2] *p* value of the CAD system obtained by operator 2 versus the experienced radiologist

[3] *p* value of the CAD system obtained by operator 3 versus the experienced radiologist. The McNemar test was used for the statistical analysis

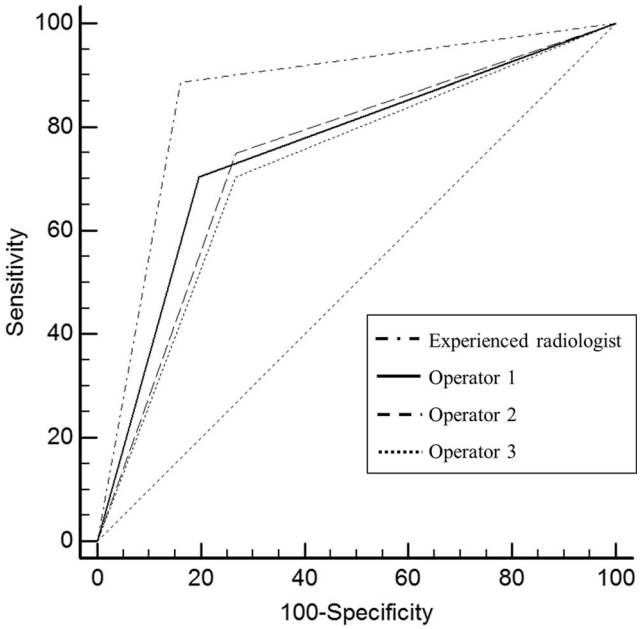

Fig. 3 A comparison of receiver operating characteristic (ROC) curves for the different operators. Area under the ROC curve (AUC) was 0.863 for the experienced radiologist and 0.718–0.754 for the less-experienced operators. There is a significant difference between AUC values of the experienced radiologist and less-experienced operators

[10]. In our study, we also found high diagnostic sensitivity (88.6%) and accuracy (86.0%) of the CAD system that was not statistically different from that of an experienced radiologist (84.1% and 91.0%, respectively). On the other hand, the specificity was significantly higher of the experienced radiologist (96.4%) than that of the CAD system (83.9%). Therefore, we conclude that the CAD system may be useful as a decision-making support to rule out cancer and ultimately to avoid unnecessary FNA.

However, although these results demonstrate the potential usefulness of the CAD system, they may not be generalizable since they were obtained only by experienced radiologists. Therefore, the diagnostic performance of the CAD system with less-experienced operators should be validated for its application. Here, we found that the diagnostic sensitivity and accuracy of the CAD system for less-experienced operators (70.5–75.0% and 72.0–76.0%, respectively) were significantly lower than those for the experienced radiologist. This indicates that the diagnostic performance of the CAD system depends on the operator's experience. Therefore, although the CAD system may be useful as a support for decision-making when used by an experienced radiologist, its data cannot be interpreted directly by less-experienced operators, and the interobserver variability of the US also influences the performance of the CAD system. However, regarding the extent of interobserver agreement, it was good ($\kappa = 0.658$) in this study. Kim et al reported that the interobserver variability in US interpretations of final category among nine observers (five faculty members and four residents) was poor for the residents ($\kappa = 0.11$–0.17), moderate for the faculty ($\kappa = 0.55$), and fair for all nine observers ($\kappa = 0.30$) [7]. Choi et al similarly reported that the interobserver agreement of the final assessment among the four radiologists with more than 5 years of experience was fair ($\kappa = 0.54$) [3]; those values were lower than in our study. Therefore, we postulate that the CAD system may help to increase interobserver agreement of US interpretation to a certain degree, especially for less-experienced operators; however, variability still remains and needs improvement.

Regarding the interobserver agreement for the US characteristics, there was good agreement for most of the US descriptors, except for the margin and composition. The margin determination by the semi-automated CAD system may be a factor of operator-dependency. The operator selects from among three or four candidate diagnoses presented by the CAD system and the operator's experience influences the final diagnosis. As these semi-automated steps require experience of the operators, automated process for deciding the nodule margin seems to be necessary for improving the operator dependency. Limitations remain regarding the interpretation of compositions by the CAD system: solid nodules with a marked hypoechoic or hypoechoic component were commonly misinterpreted as a partially cystic or partially solid nodule in our study. Further validation pertaining to this issue and an assessment based on a larger study are required to improve the current CAD system.

There were several limitations to our study. First, our sample size was small (100 thyroid nodules in 76 patients). We included only thyroid nodules > 1.0 cm that had undergone US-guided FNA, which was usually performed when suspicious US features were noted or on the largest nodule when no suspicious feature was detected. Therefore, selection bias may have been present. Second, we included only thyroid nodules with final diagnoses confirmed by FNA or a surgical specimen, which leads to a higher-than-average nodule malignancy rate. In addition, since our institution was a referral center dealing

Table 4 Interobserver agreement for the CAD system obtained with four different operators

US criteria	Observers ($n = 4$)
Composition	0.444
Echogenicity	0.723
Orientation	0.710
Margin	0.479
Spongiform	0.797
Shape	–
Final diagnosis	0.658

The Fleiss multirater kappa statistic was used to obtain an overall κ value among multiple observers

US ultrasonography

with patients with more serious disease, a higher malignancy rate may affect the diagnostic performances of the CAD system. Third, we included only nodules > 1.0 cm to enable clear CAD diagnoses, which might have influenced the diagnostic performance of the CAD system. Fourth, a high percentage of PTCs (97.4%) with a relatively low percentage of FTCs and FVPTCs may influence the diagnostic performances of the CAD system and the radiologist. Fifth, the CAD system does not yet evaluate calcification. Further technical developments will improve the performance of the CAD system.

In conclusion, the CAD system may have a potential decision-making support in the thyroid cancer diagnosis. However, operator dependency still remains and needs improvement.

Funding The authors state that this work was supported by the National Research Foundation of Korea (# 2017R1C1B5016217).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Eun Ju Ha.

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

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was obtained from all patients before they underwent US.

Ethical approval This study was approved by our institutional review board.

Methodology
• Prospective case-control study

References

1. Haugen BR (2017) 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: what is new and what has changed? Cancer 123:372–381
2. Shin JH, Baek JH, Chung J et al (2016) Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol 17:370–395
3. Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ (2010) Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid 20:167–172
4. Kim HG, Kwak JY, Kim EK, Choi SH, Moon HJ (2012) Man to man training: can it help improve the diagnostic performances and interobserver variabilities of thyroid ultrasonography in residents? Eur J Radiol 81:e352–e356
5. Park CS, Kim SH, Jung SL et al (2010) Observer variability in the sonographic evaluation of thyroid nodules. J Clin Ultrasound 38: 287–293
6. Park SH, Kim SJ, Kim EK, Kim MJ, Son EJ, Kwak JY (2009) Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules. AJR Am J Roentgenol 193:W416–W423
7. Kim SH, Park CS, Jung SL et al (2010) Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules. Korean J Radiol 11:149–155
8. Acharya UR, Sree SV, Krishnan MM et al (2014) Computer-aided diagnostic system for detection of Hashimoto thyroiditis on ultrasound images from a Polish population. J Ultrasound Med 33:245–253
9. Chang Y, Paul AK, Kim N et al (2016) Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. Med Phys 43:554
10. Choi YJ, Baek JH, Park HS et al (2017) A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. Thyroid 27:546–552
11. Li LN, Ouyang JH, Chen HL, Liu DY (2012) A computer aided diagnosis system for thyroid disease using extreme learning machine. J Med Syst 36:3327–3337
12. Lim KJ, Choi CS, Yoon DY et al (2008) Computer-aided diagnosis for the differentiation of malignant from benign thyroid nodules on ultrasonography. Acad Radiol 15:853–858
13. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382
14. Kundel HL, Polansky M (2003) Measurement of observer agreement. Radiology 228:303–308
15. Chen KY, Chen CN, Wu MH et al (2014) Computerized quantification of ultrasonic heterogeneity in thyroid nodules. Ultrasound Med Biol 40:2581–2589