

Observer Variability and the Performance between Faculties and Residents: US Criteria for Benign and Malignant Thyroid Nodules

Sung Hun Kim, MD¹
Chang Suk Park, MD¹
So Lyung Jung, MD¹
Bong Joo Kang, MD¹
Jee Young Kim, MD¹
Jae Jung Choi, MD¹
Ye Il Kim, MD¹
Jin Kyung Oh, MD¹
Jung Suk Oh, MD¹
Hanna Kim, MD¹
Seung Hee Jeong, MD²
Hyeon Woo Yim, MD²

Index terms :

Thyroid nodule
Ultrasonography
Observer variation
Faculty
Internship and residency

DOI:10.3348/kjr.2010.11.2.149

Korean J Radiol 2010; 11: 149-155

Received August 28, 2009; accepted after revision November 18, 2009.

¹Department of Radiology, College of Medicine, The Catholic University of Korea, Seoul 137-701, Korea;

²Department of Preventive Medicine, College of Medicine, The Catholic University of Korea, Seoul 137-701, Korea

Address reprint requests to:

Chang Suk Park, MD, Department of Radiology, Incheon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, #665, Bupyeong-6dong, Bupyeong-gu, Incheon 403-720, Korea.

Tel. (8232) 510-5898

Fax. (8232) 529-0964

e-mail: blounse@catholic.ac.kr

Objective: To evaluate the interobserver variability and performance in the interpretation of ultrasonographic (US) findings of thyroid nodules.

Materials and Methods: 72 malignant nodules and 61 benign nodules were enrolled as part of this study. Five faculty radiologists and four residents independently performed a retrospective analysis of the US images. The observers received one training session after the first interpretation and then performed a secondary interpretation. Agreement was analyzed by Cohen's kappa statistic. Degree of performance was analyzed using receiver operating characteristic (ROC) curves.

Results: Agreement between the faculties was fair-to-good for all criteria; however, between residents, agreement was poor-to-fair. The area under the ROC curves was 0.72, 0.62, and 0.60 for the faculties, senior residents, and junior residents, respectively. There was a significant difference in performance between the faculties and the residents ($p < 0.05$). There was a significant increase in the agreement for some criteria in the faculties and the senior residents after the training session, but no significant increase in the junior residents.

Conclusion: Independent reporting of thyroid US performed by residents is undesirable. A continuous and specialized resident training is essential to enhance the degree of agreement and performance.

The thyroid nodule is a disease entity that is commonly encountered in a clinical setting. It can be palpated in 4–8% of general adults, and its detection rate ranges from 10–41% on ultrasonography (US) (1). At the present time, thyroid US is the most accurate imaging modality in rendering a diagnosis of a thyroid nodule. Many studies have been conducted to examine the US findings that differentiate benign and malignant nodules (2–6). On fine needle aspiration cytology (FNAC), which is performed based on the US findings, up to 9–15% of the nodules have been reported to be malignant (7, 8). Radiologists play a crucial role in detecting lesions on thyroid US and determining whether a tissue biopsy is needed. However, US diagnostic accuracy varies depending on the investigator. At most hospitals, an US is performed by experienced radiologists. At resident-training hospitals, however, it is performed by radiology residents and faculty radiologists.

In this study, we evaluated the discrepancy in the US interpretations of thyroid nodules between faculties and the radiology residents. We also determined whether interobserver agreement would be enhanced at the second interpretation following a training session.

MATERIALS AND METHODS

Patients

Institutional Review Board approval was obtained for this retrospective study, and informed consent requirement was waived. A total of 133 cases from 108 patients that were between January and December of 2007 were included in this study. The patient population was made up of 72 malignant nodule cases that were surgically confirmed, as well as 61 cases of benign nodules that were confirmed by surgery or by FNAC. Our patients consisted of 96 women (89%) and 12 men (11%) and the mean age was 64.5 years (range 26–74 years). The 72 cases of malignant nodules consisted of one case of follicular adenocarcinoma, three cases of medullary carcinoma, and 68 cases of papillary carcinoma. The benign nodules included 53 cases that were negative for malignancy by two sessions of FNAC at a 6-month interval, and eight cases that were surgically confirmed (four cases of nodular hyperplasia and four cases of follicular adenoma). The mean size of the malignant nodules was 11.9 mm (range 4.4–27.8 mm), whereas the size of the benign nodules was 25.7 mm (4.8–45.6 mm).

Study Design

This study was designed to measure the interobserver agreement and performance discrepancy between the faculties and radiology residents through the first interpretation, and to determine whether these parameters would be improved by a training session.

At our institution, all residents undergo a training schedule, which includes six months of neuroradiology imaging (three months in 2nd year resident, two months in 3rd year resident and one month in 4th year resident), including performing the thyroid US and FNAC and reporting it using the US criteria outlined by the Thyroid Study Group of the Korean Society of Neuroradiology and Head and Neck Radiology (KSNHNR) (6, 9).

To evaluate the observer variability and performance discrepancy according to years of experience, five faculties with 2, 4, 4, 5, and 8 years of thyroid imaging experience, respectively, and four residents reviewed the imaging without a special training session. To evaluate the difference of the agreement and performance between senior and junior resident groups, four residents were chosen; two second-year residents who had completed the two months of thyroid imaging training and two fourth-year residents who had completed the six months of training.

To examine the effect of training on the interobserver agreement, the observers were given one training time, at eight weeks after the first interpretation. In the training

session, the US criteria for thyroid nodules were reviewed, and cases were discussed. A baseline consensus in the lexicon was reached through the discussion for 20 cases of pathologically confirmed benign and malignant nodules that were not included in the current study. Patient clinical histories, previous imaging results, and pathologic results were not available to the nine observers during the discussion. After this training session, the primary data were randomly arranged, and then the secondary data were reviewed independently.

Image Analysis

The US images were acquired using a 5–12 MHz linear probe (HDI 5000, Advanced Technology Laboratories; Bothell, WA). FNAC was performed by two faculty radiologists (4 and 3 years of experience, respectively).

One investigator selected two representative transverse and longitudinal US images for each lesion and saved the images to PowerPoint (Microsoft Corporation, Redmond, WA). The images were reviewed by five faculty radiologists and four radiology residents. Each observer independently analyzed the images and were unaware of the clinical information and pathologic results.

Each nodule was analyzed based on the US criteria that were proposed by the Thyroid Study Group of the Korean Society of Neuroradiology and Head and Neck Radiology (KSNHNR).

The internal contents of the nodules were defined as solid (more than 90% solid component), predominantly solid (more than 50% solid component, but less than 90%), predominantly cystic (more than 10% solid component, but less than 50%), and cystic (less than 10% solid component). The shape of the nodules was defined as ovoid to round, taller than wide (the anteroposterior dimension longer than the transverse diameter), or irregular (neither round to oval nor taller than wide). The border characteristics were defined as well-defined smooth, spiculated, or ill defined. The echogenicity was defined as markedly hypoechoic (lower than the echogenicity of the strap muscle), hypoechoic (lower than the echogenicity of the thyroid gland), isoechoic (equal to the thyroid echogenicity), or hyperechoic (higher than the echogenicity of the thyroid gland). The calcifications were defined as microcalcifications (tiny punctate hyperechogenicities either with or without acoustic shadowings less than 1 mm in diameter), macrocalcifications (larger than 1 mm in diameter), or rim calcifications. The malignant US characteristics were defined as a shape that was taller than wide, with a spiculated border, marked hypoechoic, and micro- or macro-calcifications. If a nodule had any feature that was consistent with malignancy, it was classified as a

suspicious malignant nodule. A probably benign nodule was classified as either completely cystic or predominantly cystic with a comet tail artifact. An indeterminate nodule was defined as a nodule that was not suspicious for malignancy or probably benign findings.

Statistical Analysis

All kappa statistics were calculated using SAS version 8.0 (MAGREE SAS Macro program, Cary, NC) to assess the proportion of interobserver agreement. The method for estimating an overall kappa value in the case of multiple observers and multiple categories is based on the work of Fleiss (10). The Kappa value represents the degree of agreement in excess of that expected by chance. It is a real number ranging from 0 to 1, and the greater the value, the higher the level of agreement. The level of agreement for Cohen’s kappa is usually defined as follows: less than 0.20 as poor agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as good agreement, and greater than 0.80 as excellent agreement (11, 12).

For the evaluation of the overall observer performance, receiver operating characteristic (ROC) curves for each observer were obtained by the MedCalc version 10.1.6.0 (MedCalc Software, Ghent, Belgium). Differences in the areas under the ROC curves were assessed with a univariate z-score test. The sensitivity, specificity, positive predictive value, and negative predictive value for each category were determined for a subgroup of the observers. A *p*-value of less than 0.05 was considered to indicate a statistically significant difference.

RESULTS

Observer Agreement

The interobserver variability for the nine observers and each subgroup (faculties, residents, senior residents and junior residents) is represented in Table 1. In all of the nine observers, there was a moderate degree of interobserver agreement for calcification ($\kappa = 0.44$), there was a fair degree of agreement for margin, internal content, category,

Table 1. Interobserver Agreement

US Criteria	Primary / Secondary Interpretation				
	Observer (n = 9)	Faculty (n = 5)	Resident (n = 4)	Senior Resident (n = 2)	Junior Resident (n = 2)
Internal content	0.29 / 0.36	0.64 / 0.63	0.17 / 0.18	0.41 / 0.71	0.32 / 0.15
Shape	0.36 / 0.40	0.42 / 0.46	0.29 / 0.34	0.24 / 0.44	0.26 / 0.21
Margin	0.22 / 0.25	0.34 / 0.40	0.19 / 0.16	0.16 / 0.12	0.25 / 0.18
Echotexture	0.19 / 0.28	0.26 / 0.45	0.10 / 0.17	0.15 / 0.22	0.06 / 0.05
Echogenicity	0.37 / 0.46	0.57 / 0.56	0.23 / 0.34	0.48 / 0.61	0.12 / 0.20
Calcification	0.44 / 0.47	0.55 / 0.63	0.42 / 0.42	0.61 / 0.57	0.27 / 0.44
Category	0.30 / 0.41	0.55 / 0.55	0.11 / 0.23	0.17 / 0.43	0.17 / 0.10

Note.— In lexicons written in bold character, degree of agreement was higher at secondary interpretation, which was performed following training session (*p* < 0.05).

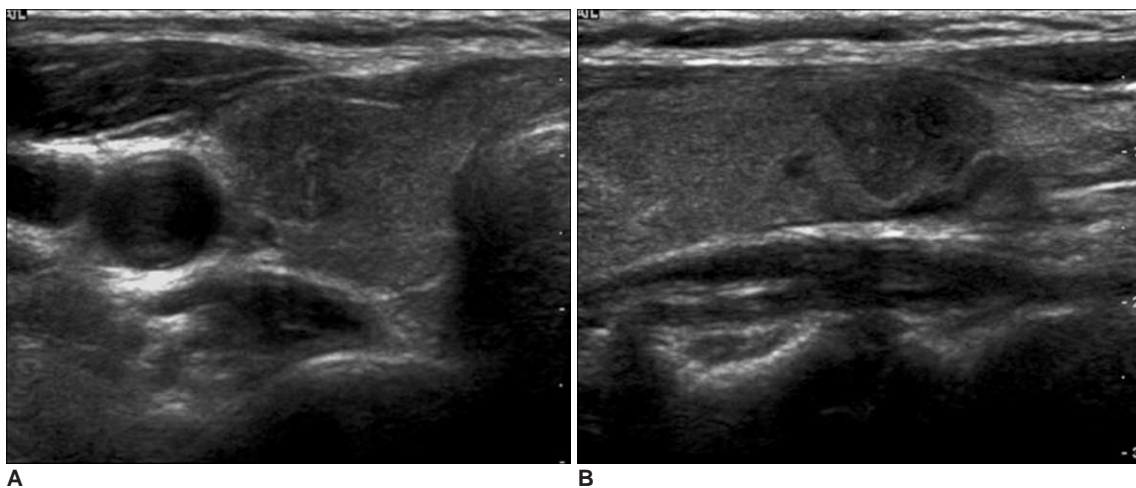


Fig. 1. Right papillary carcinoma in 47-year-old woman. Axial (A) and sagittal (B) images representing case that showed moderate interobserver agreement for presence of microcalcifications. This case was described as having microcalcifications by all faculty radiologists and three out of four residents. One resident described it as having macrocalcifications.

shape, and echogenicity ($\kappa = 0.22, 0.29, 0.30, 0.36,$ and $0.37,$ respectively), and there was a poor degree of agreement for echotexture ($\kappa = 0.19$) (Figs. 1–3). For the five faculties, the interobserver agreement was good for internal content ($\kappa = 0.64$), there was a moderate degree of agreement for shape, echogenicity, calcification, and category ($\kappa = 0.42, 0.57, 0.55,$ and $0.55,$ respectively), and there was a fair degree of agreement for margin and echotexture ($\kappa = 0.34$ and $0.26,$ respectively). For the four residents, the interobserver agreement was moderate for calcification ($\kappa = 0.42$) and fair for echogenicity and shape ($\kappa = 0.23$ and $0.29,$ respectively). For the remaining criteria, a poor degree of agreement was observed.

The agreement for calcification was higher in both the

residents and the faculties (Fig. 1), and the agreement for echotexture was lower in both groups (Fig. 2). The agreement was one level higher for internal content, echogenicity, and calcification in the senior residents in relation to the junior residents. The agreement for margin was one level higher in the junior residents, but the degree of agreement was fair. There were similar degrees of agreement for shape, echotexture, and category between the two resident groups, but the degree of agreement was lower than fair.

The observers were gathered to receive education on the US criteria, including a consensus interpretation, and then performed the secondary interpretation. The secondary interpretation demonstrated that all of the nine observers

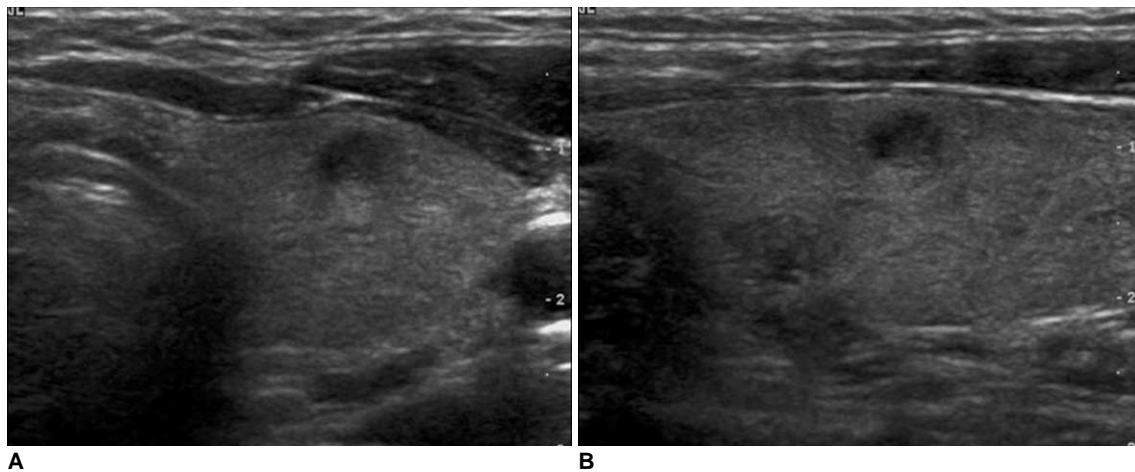


Fig. 2. Left papillary carcinoma in 42-year-old man. Axial (A) and sagittal (B) images representing case that showed poor interobserver agreement for echotexture. At primary interpretation, three of five faculties and two of four residents described this case as having heterogeneous echotexture. Remaining observers described carcinoma as being homogeneous. At secondary interpretation, all faculties agreed on heterogeneous echotexture. In residents, however, no changes were seen.

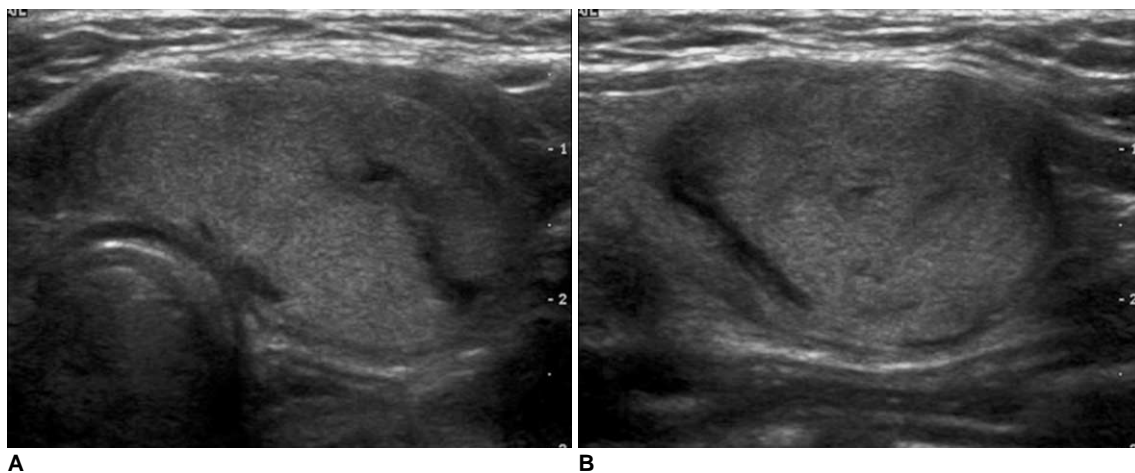


Fig. 3. Left follicular adenoma in 18-year-old woman. Axial (A) and sagittal (B) images representing case that showed fair interobserver agreement for category. Faculty radiologists classified this case as indeterminate nodule at both primary and secondary interpretation. At primary interpretation, senior residents classified case as indeterminate ($n = 1$) and probably benign ($n = 1$). At secondary interpretation, however, they all classified case as indeterminate. Junior residents classified case as indeterminate and suspicious malignant at both primary and secondary interpretation.

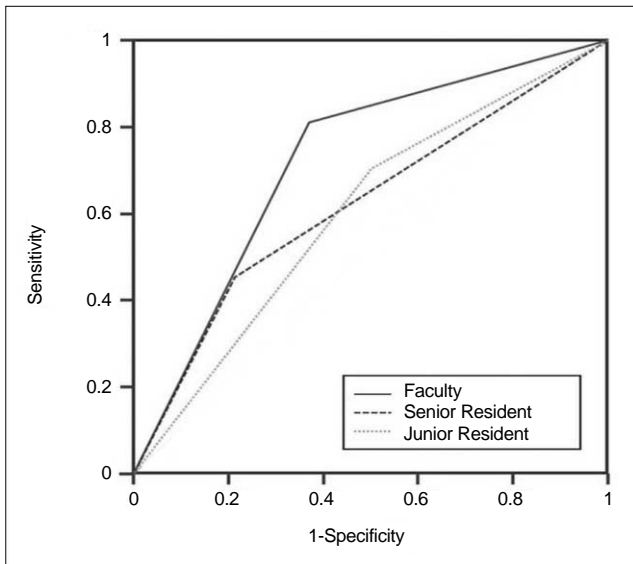


Fig. 4. Receiver operating characteristic curves for faculty, senior residents, and junior residents. Area under curves was 0.72 for faculty, 0.62 for senior residents and 0.60 for junior residents. There is significant difference between area under the receiver operating characteristic curves values of faculties and residents.

had a significantly greater degree of agreement for all of the criteria ($p < 0.05$). With regards to the changes seen in each group, the kappa value of some criteria was significantly greater in the faculty radiologists (in four criteria) and the senior residents (in two criteria), but there was no significant change for the junior residents. For the faculties, excluding criteria such as internal content, echogenicity and category, whose kappa value was the highest in the first interpretation, all of the criteria showed an increase in the degree of agreement. Particularly in echotexture, whose kappa value was the lowest, the degree of agreement increased from fair to moderate (Fig. 2). Following the training, the degree of agreement was higher than moderate, excluding the margin ($\kappa = 0.40$). In the senior residents, there was a significantly greater kappa value for internal content and category. For the junior residents; however, there was no significant increase in kappa value and there was a poor degree of agreement for all the criteria except for the calcification and the shape (Fig. 3).

Diagnostic Performance

The sensitivity and specificity obtained from the faculties, senior residents, and junior residents are summarized in Table 2. For the faculty radiologists, the sensitivity was 75%, the specificity was 65%, the positive predictive value (PPV) for a suspicious malignant nodule was 77%, and the negative predictive value (NPV) for a probably benign nodule was 69%. For the senior residents,

Table 2. Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Values

	Faculties	Senior Residents	Junior Residents
Sensitivity (%)	74.5	40.3	70.2
Specificity (%)	65.1	78.4	50.0
PPV (%)			
Suspicious malignant	76.5	74.0	62.2
Indeterminate	33.9	53.6	48.2
NPV (%)	68.9	56.2	59.6

Note.— PPV = positive predictive value, NPV = negative predictive value

the sensitivity was 40% and lower than faculty’s results, but the specificity was 78% and was higher than faculty’s results. For the junior residents, the sensitivity was 70% and the specificity was 50%, which are both lower than the faculty’s results.

The area under the receiver operating characteristic curves (AUC) was 0.72 for the faculty radiologists, 0.62 for the senior residents and 0.60 for the junior residents (Fig. 4). The performance of the faculties was better than the senior and junior residents ($p = 0.007$ and $p = 0.003$, respectively). However, for the comparison between the senior and junior residents, there was no significant difference in the diagnostic performance. Following the training session, the AUC was slightly greater for the faculty and the senior residents (0.73 and 0.65, respectively). For the junior residents, however, the AUC was lower (0.57). In each subgroup, the performance showed no significant difference compared to the primary interpretation.

DISCUSSION

The US findings have been reported to be of great help in distinguishing malignant from benign thyroid nodules (1–6). The US findings that have been reported for malignant thyroid nodules included completely solid or predominantly solid nodules, a hypoechoogenicity comparable to strap muscles, an irregular margin, intranodular microcalcifications, a taller than wide orientation, and an increase in intranodular vascularity. Moon et al. (6) reported that the findings suggested malignancy in those nodules with a taller than wide shape, a spiculated margin, a marked hypoechoogenicity, and intranodular calcifications. This study showed a high sensitivity (94%) and negative predictive value (96%), and was performed in accordance to the Thyroid Study Group of the Korean Society of Neuroradiology and Head and Neck Radiology (KSNHNR).

At most of the academic medical centers, US of the

abdomen as well as the superficial organs, including the thyroid gland and breast, is performed by radiology residents and experienced radiologists. The US performed by residents were interpreted in the preliminary form. Then, the final report was confirmed by the attending faculty. To date, the studies about the difference in agreement for image interpretation between radiology residents and experienced radiologists have been conducted with a main focus on the examinations that were performed at an emergency center, which depend on the preliminary interpretation of radiology residents after working hours (weekdays and on weekends). Previous literature has reported that the degree of agreement was greater than 90% between radiology residents and neuroradiologists in the interpretation of head CT and neuroradiology MR imaging (13–15). According to a study that examined the difference in agreement for the interpretation of body CT and US at a level I trauma center, the agreement rate between the residents and the faculty was more than 90% (16). Under the clinical suspicion of pulmonary embolism, an interobserver variability between them showed a value of $\kappa = 0.73$, indicating a good level of agreement for the interpretation of CT pulmonary angiography (17).

To our knowledge, no studies have examined the difference in the interpretation of US images between residents and faculty. In the current study, the interobserver agreement between the faculties and residents showed a moderate agreement for calcification, but showed poor agreement for echotexture. There was a fair degree of agreement for other criteria. The interobserver agreement among the five faculties was greater than for the residents. In addition, excluding a moderate agreement for calcification and a fair agreement for shape and echogenicity, the residents showed poor agreement overall.

According to a study by Moon et al. (6), the observer agreement was good or higher for internal content and microcalcifications and was moderate for shape, margin, echotexture, and echogenicity for three experienced radiologists. Compared to our results, the degree of agreement was higher for Moon et al.. Presumably, this might be because Moon et al.'s study enrolled a small number of observers and only experienced radiologists provided diagnoses. Furthermore, Moon et al. performed two sessions of training before the study and thereby achieved a baseline consensus. Wienke et al. (18) reported an interobserver variability between two experienced radiologists and showed that there was a good or higher agreement for calcification and the amount of cystic component. In addition, there was fair agreement for echogenicity and there was poor agreement for border.

Compared to the results of the five faculty radiologists, their results showed lower agreement for echogenicity and border, as well as higher agreement for calcifications. Their results were similar with ours with respect to the amount of cystic component (i.e., the internal content).

The comparison of faculties and residents suggests that diagnostic performance was significantly higher for the faculties, but there was no significant difference between the junior and senior residents. AUC was 0.72 for the faculties, 0.62 for the senior residents, and 0.60 for the junior residents.

The secondary interpretation after the training session indicated that all nine observers had a significantly greater degree of agreement for all the categories ($p < 0.05$). The faculty radiologists showed a significant increase in agreement in four categories and in two categories for the senior residents. For the junior residents, agreement was improved, but not to a significantly and there was still a poor agreement for most lexicons. The AUC was slightly greater for the faculties and senior residents, but decreased in the junior residents, after the training session. These results suggest that one training session was not of great help in improving the agreement level and performance in the junior residents. The number of categories with significantly greater agreement was superior in the faculties in relation to the senior residents. These results suggest that attending radiologists would require a more meticulous review and confirmation of the preliminary interpretation by the residents. The clinical experience through the practice of thyroid US and FNAC, the consistent feedback from the correlation with the pathologic results, and the continuous consensus interpretation would be necessary for the effective training of the residents.

There are several limitations to the current study. Namely, there is a selection bias since only images specifically selected by an investigator were used. In addition, there was a relatively high percentage of malignant nodules in the sample. Presumably, this is because our institution is a referral center for patients with suspicious thyroid lesions. The low incidence of benign nodules is related to the fact that benign cases were only included when there were negative for malignancy by two sessions of FNAC at a 6-month interval. The current study enrolled residents as two groups and it is questionable whether these residents can be representative of all residents. Also, the current study had a retrospective design. In other words, the observers were aware that their interpretations did not affect the patient care, thus lessening the concern that a patient would be recalled for a biopsy. This may produce a discrepancy from the results that would have been obtained in an actual clinical setting (19). Lastly, we

saved the images in PowerPoint for review, which could have resulted in a decrease in image resolution.

In conclusion, the observer agreement for all the criteria was higher in faculties than for residents. Moreover, diagnostic performance was significantly higher in faculties than residents. Consequently, we suggest that independent interpretation by residents after examining the US is undesirable. A continuous and specialized resident training program would be essential to enhance the degree of agreement and performance in residents.

References

1. Frates MC, Benson CB, Charboneau JW, Cibas ES, Clark OH, Coleman BG, et al. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. *Radiology* 2005;237:794-800
2. Kim EK, Park CS, Chung WY, Oh KK, Kim DI, Lee JT, et al. New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. *AJR Am J Roentgenol* 2002;178:687-691
3. Chan BK, Desser TS, McDougall IR, Weigel RJ, Jeffrey RB Jr. Common and uncommon sonographic features of papillary thyroid carcinoma. *J Ultrasound Med* 2003;22:1083-1090
4. Iannuccilli JD, Cronan JJ, Monchik JM. Risk for malignancy of thyroid nodules as assessed by sonographic criteria: the need for biopsy. *J Ultrasound Med* 2004;23:1455-1464
5. Shimura H, Haraguchi K, Hiejima Y, Fukunari N, Fujimoto Y, Katagiri M, et al. Distinct diagnostic criteria for ultrasonographic examination of papillary thyroid carcinoma: a multicenter study. *Thyroid* 2005;15:251-258
6. Moon WJ, Jung SL, Lee JH, Na DG, Baek JH, Lee YH, et al. Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. *Radiology* 2008;247:762-770
7. Nam-Goong IS, Kim HY, Gong G, Lee HK, Hong SJ, Kim WB, et al. Ultrasonography-guided fine-needle aspiration of thyroid incidentaloma: correlation with pathological findings. *Clin Endocrinol (Oxf)* 2004;60:21-28
8. Frates MC, Benson CB, Doubilet PM, Kunreuther E, Contreras M, Cibas ES, et al. Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. *J Clin Endocrinol Metab* 2006;91:3411-3417
9. Thyroid Study Group of the Korean Society of Neuroradiology and Head and Neck Radiology (KSNRHNR). *Thyroid gland: imaging diagnosis and intervention*, 1st ed. Seoul: Ilchokak, 2008
10. Fleiss JL. *Statistical methods for rates and proportions*, 2nd ed. New York: John Wiley & Sons Inc., 1981
11. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174
13. Wysoki MG, Nassar CJ, Koenigsberg RA, Novelline RA, Faro SH, Faerber EN. Head trauma: CT scan interpretation by radiology residents versus staff radiologists. *Radiology* 1998;208:125-128
14. Erly WK, Berger WG, Krupinski E, Seeger JF, Guisto JA. Radiology resident evaluation of head CT scan orders in the emergency department. *AJNR Am J Neuroradiol* 2002;23:103-107
15. Filippi CG, Schneider B, Burbank HN, Alsofrom GF, Linnell G, Ratkovits B. Discrepancy rates of radiology resident interpretations of on-call neuroradiology MR imaging studies. *Radiology* 2008;249:972-979
16. Carney E, Kempf J, DeCarvalho V, Yudd A, Noshier J. Preliminary interpretations of after-hours CT and sonography by radiology residents versus final interpretations by body imaging radiologists at a level 1 trauma center. *AJR Am J Roentgenol* 2003;181:367-373
17. Yavas US, Calisir C, Ozkan IR. The interobserver agreement between residents and experienced radiologists for detecting pulmonary embolism and DVT with using CT pulmonary angiography and indirect CT venography. *Korean J Radiol* 2008;9:498-502
18. Wienke JR, Chong WK, Fielding JR, Zou KH, Mittelstaedt CA. Sonographic features of benign thyroid nodules: interobserver reliability and overlap with malignancy. *J Ultrasound Med* 2003;22:1027-1031
19. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;249:47-53