Ultrasound elastography for assessing thyroid nodules: Interrater variability and diagnostic performance

Jieun Koh

Department of Medicine

The Graduate School, Yonsei University

Ultrasound elastography for assessing thyroid nodules: Interrater variability and diagnostic performance

Directed by Professor Jin Young Kwak

The Master's Thesis
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree
of Master of Medical Science

Jieun Koh

June 2014

This certifies that the Master's Thesis of
Jieun Koh is approved.

Thesis Supervisor : Jin Young Kwak

Thesis Committee Member#1 : Hee Jung Moon

Thesis Committee Member#2 : Jeong Seon Park

The Graduate School
Yonsei University

June 2014

ACKNOWLEDGEMENTS

I acknowledge my deep gratitude to Professor Jin Young Kwak, who is my thesis director, for supporting my efforts with total commitment and facilitating every step of the process. My appreciation for his guidance and encouragement is tremendous. I am also indebted to Professor Hee Jung Moon and Jeong Seon Park, for their help for pertinent advice to assure the superior quality of this paper.

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Ultrasound elastography for assessing thyroid nodules: Interrater variability and diagnostic performance

Jieun Koh

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Jin Young Kwak)

Purpose: To validate interrater variability for strain ultrasound elastography (USE) and compare diagnostic performances of a combination of gray scale ultrasound (US) with USE with gray-scale US. **Methods:** Three raters from different institutions evaluated gray-scale US images and USE video files of 443 cytopathologically proven benign or malignant thyroid nodules during a 3 month term. Interrater variability of gray-scale US and USE using the Asteria or Rago criteria were evaluated. We compared diagnostic performances for predicting malignancy on gray-scale US with the combination of gray-scale US and USE for each rater. **Results:** Interrater variability was not statistically different between USE using the Asteria criteria and gray-scale US, however USE using the Rago criteria demonstrated the lowest interrater agreement (P<0.043). Sensitivity was increased in all three raters by adding USE to gray-scale US (81.3-88.3%, 75.4-85.4%) compared to gray-scale US (70.4-80.8%) alone. Specificity were decreased by adding USE to gray-scale US (51.7-59.1, 59.1-73.9%) compared to gray-scale US alone (69.0-82.8%). **Conclusions:**

USE showed comparative interrater variability to gray-scale US. However when USE is added to gray-scale US, additional diagnostic yield is limited compared to gray-scale US alone.

Key words : elastography; thyroid nodule; interrater variability; ultrasound

Ultrasound elastography for assessing thyroid nodules: Interrater variability and diagnostic performance

Jieun Koh

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Jin Young Kwak)

I. INTRODUCTION

Gray scale ultrasound (US) is the most sensitive test to detect thyroid lesions; however the differentiation of benign and malignant nodules is not highly accurate,[1] and its diagnostic value varies considerably from study to study.[2-6] Ultrasound elastography (USE) enables assessment of tissue consistency by differentiating hard from soft nodules, and it supplements the diagnostic limitations of gray-scale US.[7-20] Previous studies suggested improved or comparative diagnostic performances of USE compared with gray-scale US in differentiating benign and malignant thyroid nodules.[8-10,13,14,16] Diagnostic performances have also been improved with a combination of gray-scale US and USE.[19] Contrary to these positive results, several studies have failed to prove the superiority of USE compared with gray-scale US.[12,15,17,21] Moreover, the combination of USE and gray-scale US was inferior to gray-scale US in

some cases.[15]

The other technical issue with USE besides variable diagnostic performance is limited interrater agreement, an issue first reported by Park et al. using strain USE, however without subjective monitoring methods for compression.[22] Other consecutive studies reported increased interrater agreement using subjective monitoring methods for compression on strain USE.[23-25] Shear wave USE showing fair to excellent reproducibility in neck lesions including thyroid nodules tend to have higher interrater agreement compared with strain USE.[26-28] So far, studies to evaluate the interrater variability of strain USE have been limited to small sample sizes and only performed by raters from the same institution.[22-25] Therefore, we investigated to validate the interrater agreement for strain USE as well as gray-scale US in a relatively large number of thyroid nodules by three radiologists from different institution and compared diagnostic performances of gray-scale US with a combination of gray-scale US with USE.

II. MATERIALS AND METHODS

1. Patients

The institutional review board approved this retrospective study and required neither patient approval nor informed consent for our review of patients' images and records. From November 2011 to January 2012, 583 nodules in 465 consecutive patients underwent FNA or staging US with strain USE. We

excluded nodules measured less than 5 mm or equal to or larger than 30 mm (n=65) and nodules with cytologic results of suspicious malignant, atypia, and nondiagnostic (n=61) with no further surgical intervention. Among 457 nodules, 194 were pathologically confirmed by surgery, and 263 were cytologically proved to be benign or malignant nodules with no further surgical intervention. Among them, 14 nodules were excluded due to the poor quality of USE video files or gray-scale US images. Finally, 443 nodules in 426 patients were included in this study, and among them 17 patients had two nodules. The mean age of the patients was 47 ± 12, and 361 were female, and 82 were male. The mean size of the nodules was 11 ± 5.6 mm, and 212 nodules were equal to or less than 10mm, and the rest were larger than 10mm.

2. Gray-scale US and USE

Gray-scale US was performed initially with a 6–14-MHz linear array transducer (EUB-7500; Hitachi Medical, Tokyo, Japan) by radiologists with 1 to 15 years of experience in thyroid imaging. Transverse and longitudinal images of thyroid nodules were captured and stored for later image interpretation. After performing gray-scale US, USE was performed by the same radiologist with the same US unit. All USE images were obtained in longitudinal planes with the freehand technique. Each radiologist had at least two months of experience with the machine, and had performed USE for more than 100 nodules during training.

The probe was positioned perpendicular to the skin, and repetitive compression was applied above the targeted thyroid nodules during USE. A square region of interest was placed at the target nodule with the superior margin including subcutaneous fat and the inferior margin including the longus colli muscle. Color homogeneity within the region and pressure indicator (ranging 2-3) was monitored for optimal image acquisition. On a split-screen mode, gray-scale US images were displayed on the right while USE images that were superimposed on the corresponding gray-scale US images were displayed on the left. USE images were displayed with 256 specific colors for each pixel ranging from a color spectrum of red to blue. The softest composition was displayed in red and the hardest composition in blue.[15] USE images were obtained as video files with more than 5 seconds continuous length.

3. Image interpretation

Stored gray scale image and USE video files were reviewed by one radiology resident (J.E.K). On reviewing gray-scale US images in PACS (the picture archiving and communication system), an appropriate transverse and longitudinal view of each nodule were manually captured. All clinical data from images were removed. Images of each nodule were assigned with randomized numerical numbers and ordered. Video files less than 5 seconds in length were excluded during USE video file review. USE video files of each nodule were

also randomized with other numerical orders different from the gray-scale US images.

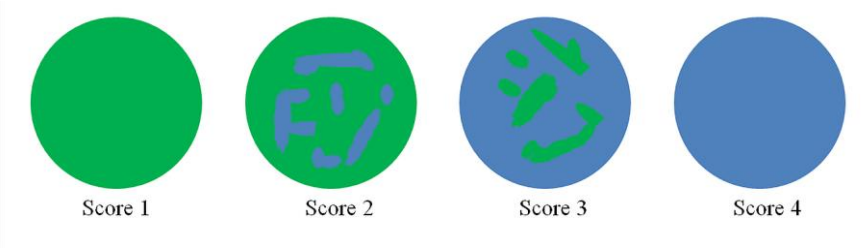
Three radiologists from different hospitals evaluated gray-scale US images and USE video files. The first radiologist (H.J.M) had 11 years of thyroid US experience and 5 years of USE experience. The second radiologist (J.S.P) had 12 years of thyroid US experience and 8 years of USE experience. The third radiologist (S.J.K) had 7 years of thyroid US experience and 5 years of USE experience. All three raters were unaware of clinical data or cytologic results. First, gray-scale US images were sent to each rater for evaluation, and interpreted results were recorded in a form collected right after image review. Three months after the gray-scale US image review, a set of USE video files was sent to each rater and the results were then recorded in another form and collected.

On the gray-scale US image interpretation form, five features of the thyroid nodules were recorded. Internal composition of nodules was recorded as solid, < 50% of cystic portion, ≥ 50% of cystic portion, and cyst. Echogenicity of nodules was recorded as hyper-, iso-, hypo-, and marked hypoechogenicity. Margin of nodules was evaluated as well circumscribed, microlobulated, or irregular. Presence of calcification in nodules was recorded as microcalcification, macrocalcification (including egg shell calcification), mixed micro- and macrocalcification, and no calcification. Shape of nodules was

interpreted as taller than wider or wider than taller. Features of malignant thyroid nodules included solid internal composition, marked hypoechogenicity, microlobulated or irregular margin, presence of microcalcification, and taller than wider shape. Final assessment was recorded based upon the presence of malignant features, with assessments as probably benign when there were no malignant features and as suspicious when one or more malignant features were present. There were 37 thyroid nodules containing macrocalcifications and two thyroid nodules had predominant cystic portions.

USE video files were reviewed and thyroid nodules were scored according to classifications of Asteria et al.[8] and Rago et al.[9] separately (Figure 1). Asteria et al. scored elasticity (the Asteria criteria) with 4 scales, and as the scale increased, elasticity decreased. Scores of 3 and 4 were classified as suspicious malignancy, and 1 and 2 as probably benign. The scoring system of elasticity by Rago et al. (the Rago criteria) used 5 scales (1-5), and a lower number indicated more elasticity while an increased number correlated with increased nodule hardness. Nodules with scores of 4 and 5 were classified as suspicious malignancy, whereas scores of 1 to 3 were classified as probably benign.

(a)



(b)

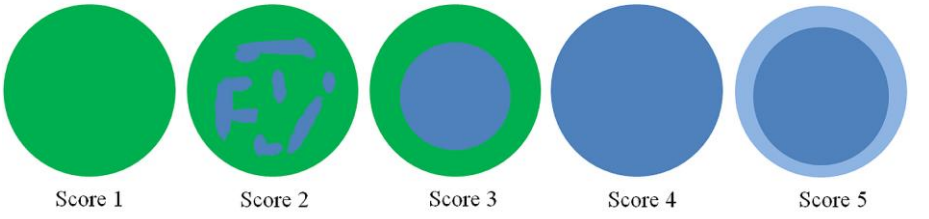


Figure 1. Schematic representation of elasticity of a thyroid nodule scored according to the Asteria criteria (a) and Rago criteria (b) a Scores from 1 to 4 indicate the following; 1 Elasticity in the whole examined area, 2 Elasticity in a large portion of the examined area, 3 No elasticity in a large portion of the examined area, 4 No elasticity in the whole examined area b Scores from 1 to 5 indicated as follows; 1 Elasticity in the whole nodule, 2 Elasticity in a large part of the nodule, 3 Elasticity only at the peripheral part of the nodule, 4 No elasticity in the nodule, 5 No elasticity in the nodule and in the posterior shadowing.

4. US-guided FNA

US-guided FNA was performed by the same radiologist who performed US. Aspiration was performed at least twice for each nodule using the freehand technique with a 23-gauge needle attached to 2-mL disposable plastic syringe. Obtained samples were expelled on to glass slides, smeared and placed immediately into 95 % alcohol for Papanicolaou staining. One of the five cytopathologists specializing in thyroid cytology interpreted the smeared samples. The Bethesda classifications were used in the cytology reports of thyroid aspirate samples.[29]

5. Data and Statistical analysis

We used cytopathological results as standard reference, with samples confirmed as malignant through pathology or FNA classified into the positive group, and samples confirmed as benign through pathology or FNA classified into the negative group. Continuous variables were analyzed using the student t-test, and categorical variables were analyzed by the Pearson's chi-square test. Interrater variability of gray-scale US and USE were evaluated between two raters using Cohen's kappa analysis for pairwise comparison. Generalized kappa using the Inter_Rater SAS Macro program was used for overall comparison among the three raters of gray-scale US and USE. Additionally, we compared kappa coefficients among final assessment of gray-scale US and elastography using

the Asteria and Rago criteria.[30] Relative strength of agreement associated with kappa statistics was poor ($\kappa \leq 0.2$), fair ($0.2 < \kappa \leq 0.4$), moderate ($0.4 < \kappa \leq 0.6$), substantial ($0.6 < \kappa \leq 0.8$), or good ($\kappa > 0.8$).[31] We calculated and compared the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy for predicting malignancy on gray-scale US with the combination of gray-scale US and USE for each rater with the generalized estimating equation (GEE) method. Analysis was performed using SAS 9.2 version (SAS Institute Inc., Cary, NC, USA.). Statistical significance was assumed when the *P* value was less than 0.05.

III. RESULTS

Of the total 443 nodules, 240 (54.2%) nodules were confirmed as malignancy, and the remaining 203 (45.8%) nodules as benign. Among the 180 nodules surgically confirmed as malignancy, 174 were conventional papillary thyroid carcinoma, 5 were follicular variant of papillary thyroid carcinoma, and 1 was medullary carcinoma. Among the 8 nodules surgically confirmed as benign, 4 nodules were adenomatous hyperplasia, 2 were lymphocytic thyroiditis, 1 was Hurthle cell adenoma, and 1 was cellular adenomatous hyperplasia. The mean size (9.23 ± 4.26 mm) of the malignant nodules was smaller than that of (13.39 ± 6.12 mm) the benign nodules with statistical significance ($P<0.001$). Gender and age were not associated with malignancy ($P=0.277$ and 0.074, respectively).

Interrater agreement analysis for each feature in gray-scale US and USE was performed (Table 1). Overall interrater agreement was substantial in margin and shape (κ=0.618 and 0.760), and moderate in composition, echogenicity, and calcification (κ=0.545, 0.417, and 0.592). Shape showed the highest level of interrater agreement and echogenicity showed the lowest level of interrater agreement among the three raters. In regard to final assessment of gray-scale US, overall interrater agreement was substantial (κ=0.621) and interrater variability between two raters was also substantial among the three raters (κ=0.603-0.644).

Table 1 Interrater variability of each feature and final assessment of gray-scale US and USE scored according to the Asteria and Rago criteria

	κ (Standard error)							
	Composition[1]	Echogenicity[1]	Margin[1]	Calcification[1]	Shape[1]	Final assessment[1]	Asteria[2]	Rago[3]
Overall	0.545 (0.044)	0.417 (0.066)	0.618 (0.080)	0.592 (0.045)	0.664 (0.037)	0.621 (0.028)	0.602 (0.028)	0.360 (0.050)
1 vs. 2	0.483 (0.046)	0.311 (0.058)	0.618 (0.037)	0.573 (0.046)	0.627 (0.041)	0.620 (0.037)	0.601 (0.037)	0.475 (0.050)
1 vs. 3	0.684 (0.038)	0.534 (0.054)	0.636 (0.037)	0.596 (0.045)	0.690 (0.038)	0.644 (0.036)	0.588 (0.038)	0.240 (0.040)
2 vs. 3	0.464 (0.045)	0.404 (0.068)	0.603 (0.038)	0.609 (0.046)	0.678 (0.038)	0.603 (0.037)	0.624 (0.037)	0.427 (0.050)

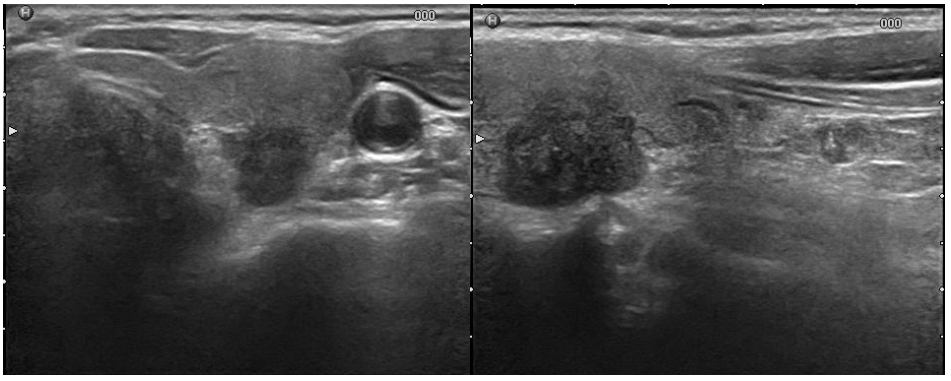
[1]Gray-scale US

[2]Asteria criteria USE scored with a 4 grade scale

[3]Rago criteria USE scored with a 5 grade designation

USE using the Rago criteria demonstrated the lowest overall and pairwise interrater agreement compared with gray-scale US or USE using the Asteria criteria (P<0.043) (Figure 2). Overall interrater agreement for USE using the Asteria criteria was substantial (κ=0.602) and for USE using the Rago criteria was fair (κ=0.360) among the three raters. Pairwise interrater agreement showed substantial agreement between rater 1 and 2 (κ=0.601) and rater 2 and 3 (κ=0.624), and moderate between 1 and 3 (κ=0.588) in USE using the Asteria criteria. In USE using the Rago criteria, interrater agreement was moderate between rater 1 and 2 (κ=0.475) and rater 2 and 3 (κ=0.427), and fair between rater 1 and 3 (κ=0.240). The lowest level of agreement was noted between rater 1 and 3 with the Rago criteria.

(a)



(b)

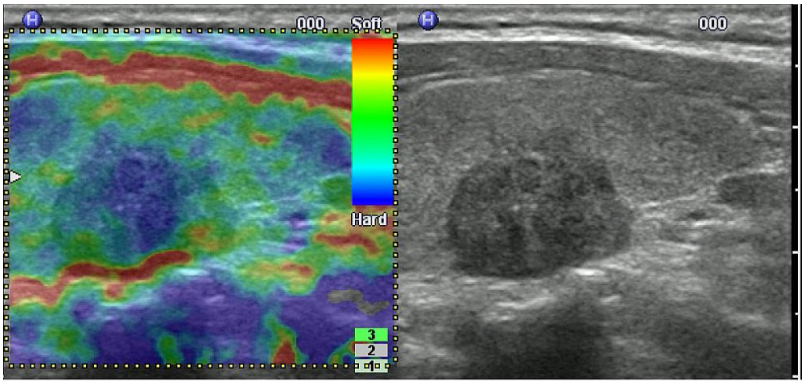


Figure 2. A 58-year-old woman with 15-mm thyroid nodule. (a) All three raters scored this nodule as suspicious on gray-scale US. (b) On USE, all raters scored this nodule as 3 by the Asteira criteria and raters 1 and 2 scored it as 2 and 3 point while rater 3 scored it as 4 point by the Rago criteria. This nodule was surgically confirmed as malignant.

Diagnostic performances were calculated for gray-scale US, USE, and the combination of gray-scale US with USE (Table 2). Sensitivity, NPV, and accuracy of gray-scale US (70.4-80.8%, 68.4-75.3%, and 72.9-76.5%, respectively) were higher than those of both USE using the Asteria criteria(45.0-59.2%, 53.2-57.8%, and 58.2-62.3%, respectively) and USE using the Rago criteria (15.4-41.3%, 49.0-53.5%, and 52.4-59.0%, respectively) in all three raters. Sensitivity, NPV, and accuracy were lower in USE using the Rago criteria compared with USE using the Asteria criteria in all three raters. Specificity was the highest in USE using the Rago criteria (79.8-96.1%) compared with USE using the Asteria criteria (63.6-73.9%) and gray-scale US (69.0-75.9%). PPV was the highest in USE using the Rago criteria (82.2%) compared with USE using the Asteria criteria (67.1%) and gray-scale US (77.5%) in rater 1. Rater 2 and 3 showed the highest PPV in gray-scale US (83.0% and 75.5%), compared with USE using the Asteria criteria (64.4% and 67.3%) and USE using the Rago criteria (73.4% and 70.7%).

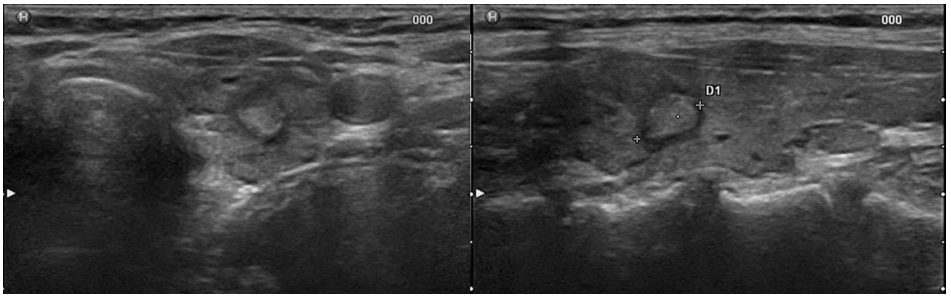
Table 2 Diagnostic performances of gray-scale US and USE scored according to the Asteria and Rago criteria, and addition of the USE using the Asteria and Rago criteria to gray-scale US

	Sensitivity(%)			Specificity(%)			PPV(%)			NPV(%)			Accuracy(%)		
Raters	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Gray-scale US	70.4	71.3	80.8	75.9	82.8	69.0	77.5	83.0	75.5	68.4	70.9	75.3	72.9	76.5	75.4
USE using the Asteria score	45.0	55.8	59.2	73.9	63.6	66.0	67.1	64.4	67.3	53.2	54.9	57.8	58.2	59.4	62.3
USE using the Rago score	15.4	28.8	41.3	96.1	87.7	79.8	82.2	73.4	70.7	49.0	51.0	53.5	52.4	55.8	58.9
Combination of gray-scale US with USE using the Asteria score	81.3	85.0	88.3	59.1	55.7	51.7	70.1	69.4	68.4	72.7	75.8	79.0	71.1	71.6	71.6
Combination of gray-scale US with USE using the Rago score	75.4	79.6	85.4	74.4	73.9	59.1	77.7	78.3	71.2	71.9	75.4	77.4	74.9	77.0	73.4

We compared diagnostic performances of gray-scale US with the combination of gray-scale US with USE (Table 2). Sensitivity was increased by adding USE using the Asteria criteria to gray-scale US (81.3-88.3%) and by adding USE using the Rago criteria to gray-scale US (75.4-85.4%) in all three raters with statistical significance compared with gray-scale US (70.4-80.8%) (Figure 3). NPV was also increased by adding USE using the Asteria criteria to gray-scale US (72.7, 75.8%) and by adding USE using the Rago criteria to gray-scale US (72.0, 75.4%) in rater 1 and 2 compared with gray-scale US (68.4, 70.9%). In rater 3, NPV was also increased by adding USE using the Asteria criteria to gray-scale US (79.0%) and by adding USE using the Rago criteria to gray-scale US (77.4%) compared with gray-scale US (75.3%), however there was no significant statistical difference (P=0.085, 0.170). Accuracy was increased by adding USE using the Rago criteria to gray-scale US (74.9%) compared with gray-scale US (72.9%) in rater 1, and in rater 2 and 3 accuracy was decreased by adding USE using the Asteria criteria to gray-scale US (71.6 and 71.6%) compared with gray-scale US (76.5 and 75.4%) with statistical significance. Specificity and PPV were decreased by adding USE using the Asteria criteria to gray-scale US (51.7-59.1% and 68.4-70.1%) and by adding USE using the Rago criteria to gray-scale US (59.1-73.9% and 71.2-28.3%) compared with gray-scale US (69.0-82.8% and 75.5-83.0%) with rater 1 being the exception; for rater 1, the addition of USE using the Rago criteria to gray-scale US showed no

statistical difference with gray-scale US alone.

(a)



(b)

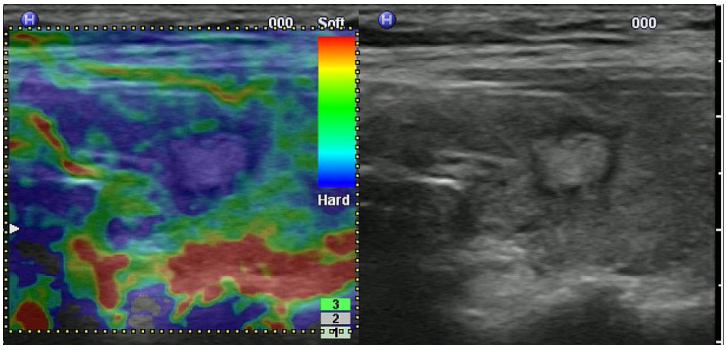


Figure 3. A 45-year-old woman with 14-mm thyroid nodule. (a) On gray-scale US, all three raters assessed this nodule as benign. b On USE, all three raters scored this nodule as 4 point by the Asteria and rater 1 and 2 both scored it as 4 point while rater 3 scored as 5 point by the Rago criteria. By fine-needle aspiration, this nodule was confirmed as benign.

IV. DISCUSSION

USE is a promising technique visualizing the elastic restoring forces of tissue that act against deformation.[7] Strain USE uses mechanically induced quasi-static shear waves and the results of tissue compression are displayed as an image.[7] Malignant nodules tend to be harder than benign lesions, and by palpation physicians can subjectively differentiate malignant nodules from benign nodules. On USE, hard lesions strain less than surrounding soft tissue, and in this manner we can differentiate malignant nodules from benign nodules objectively.[8-19]

As a complementary tool to gray-scale US, the reproducibility of USE is an important factor.[2,17,22,24] Variability can be caused throughout various USE procedures, from selection of imaging planes, compression, selection of images from dynamic sequences, to scoring.[2] There have been reports on factors affecting the poor reliability of USE and the interrater and intrarater agreement for assessing thyroid nodules including <50% green color in the region of interest box for the thyroid parenchyma, discordance in elasticity scores in the USE images, and intranodular color signal loss.[32] In this study, we evaluated interrater variability using video files of USE, excluding technical factors by a performer on USE. Among the features of gray-scale US, shape showed the highest level of agreement which was substantial, whereas echogenicity showed the lowest level of agreement which was from fair to moderate, not only

constantly between two raters but also among all three raters. Final assessment of gray-scale US showed a substantial level of agreement not only between two raters but also among the three raters. USE using the Asteria criteria showed comparable interrater variability with gray-scale US features between two raters as well as among the three raters. While USE using the Rago criteria showed less concordance than gray-scale US, USE using the Asteria criteria showed a fair to moderate level of concordance between two raters, and a fair level of concordance was noted among the three raters. This difference may result from different scales; USE using the Asteria criteria used a 4 point scoring system while USE using the Rago criteria used a 5 point scoring system and more simple definitions were applied with the Asteria criteria than the Rago criteria. A previous study using a 4 scale score system of USE which was the same as USE using the Asteria criteria showed good agreement between two raters and among three raters, which is similar result to our study.[24] Other studies showed excellent interrater agreement of USE when using a different scoring system from the system we used and compared only between two raters.[23,25]

Reported sensitivity of gray-scale US varies 83.3-94.0% and specificity varies 66-92.0% from study to study.[4,33-35] In this study, the sensitivity of USE ranged 45.0-59.1% in USE using the Asteria criteria and 15.4-41.3% in USE using the Rago criteria, and specificity ranged 66.0-73.9% in USE using the Asteria criteria and 79.8-96.1% using the Rago criteria, showing relatively low

diagnostic performances compared with initial studies (sensitivity: 94%-97%, specificity: 81%-100%) that used the same scoring system.[8,9] The diagnostic performance of USE alone was also lower compared with the final assessment of gray-scale US except for specificity in USE using the Rago criteria. However, USE using the Rago criteria showed lower sensitivities for each feature in gray-scale US. These results are different from previous studies that suggest that a high diagnostic performance is possible with USE.[10,13,14,16]

All diagnostic performances of USE using the Asteria or Rago criteria except specificity and PPV of USE using the Rago criteria were inferior to those of gray-scale US. However when combined with gray-scale US, USE elevated sensitivity and NPV, whereas specificity and PPV were decreased. USE using the Asteria criteria showed better sensitivity when added to gray-scale US than USE using the Rago criteria, however when USE using the Rago criteria was added to gray-scale US, specificity was less decreased.[14,24] These findings were comparative with other studies that also suggested increased sensitivity and decreased specificity when USE was added to gray-scale US.[14,15] Accuracy was increased compared to gray-scale US in one rater when USE using the Rago criteria was added to gray-scale US with statistical significance (P=0.019), and for the other two raters, accuracy was decreased compared to gray-scale US when USE using the Asteria criteria was added to gray-scale US (*P*=0.019).

We acknowledge that there are several limitations in this study. First, only 8

benign nodules were surgically confirmed with the remaining benign nodules being confirmed by cytologic results, and false-negative results may have existed. Second, we included thyroid nodules with macrocalcifications and cystic portions to evaluate interrater variability, another factor which may affect the diagnostic performances of USE. However, as these nodules only comprised a small portion of this study, interrater variability due to macrocalcifications or cystic portions may have had little effect on the results.[36,37] Third, although we used USE using video files, differences might occur with real time image evaluation. Fourth, radiologists with variable experiences performed elastography. This may influence variability during image acquisition.

V. CONCLUSION

In conclusion, USE using the Asteria criteria showed comparative interrater variability to gray-scale US. However when USE is added to gray-scale US, the additional diagnostic yield is limited when compared with gray-scale US alone.

REFERENCES

1. Takashima S, Fukuda H, Nomura N, Kishimoto H, Kim T, Kobayashi T. Thyroid nodules: re-evaluation with ultrasound. Journal of clinical ultrasound : JCU 1995;23:179-84.
2. Lim DJ, Luo S, Kim MH, Ko SH, Kim Y. Interobserver agreement and intraobserver reproducibility in thyroid ultrasound elastography. AJR. American journal of roentgenology 2012;198:896-901.
3. Fish SA, Langer JE, Mandel SJ. Sonographic imaging of thyroid nodules and cervical lymph nodes. Endocrinology and metabolism clinics of North America 2008;37:401-17, ix.
4. 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;178:687-91.
5. 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;91:3411-7.
6. Kovacevic DO, Skurla MS. Sonographic diagnosis of thyroid nodules: correlation with the results of sonographically guided fine-needle aspiration biopsy. Journal of clinical ultrasound : JCU 2007;35:63-7.

7. Bamber J, Cosgrove D, Dietrich CF, Fromageau J, Bojunga J, Calliada F, et al. EFSUMB guidelines and recommendations on the clinical use of ultrasound elastography. Part 1: Basic principles and technology. Ultraschall in der Medizin 2013;34:169-84.

8. Asteria C, Giovanardi A, Pizzocaro A, Cozzaglio L, Morabito A, Somalvico F, et al. US-elastography in the differential diagnosis of benign and malignant thyroid nodules. Thyroid : official journal of the American Thyroid Association 2008;18:523-31.

9. Rago T, Santini F, Scutari M, Pinchera A, Vitti P. Elastography: new developments in ultrasound for predicting malignancy in thyroid nodules. The Journal of clinical endocrinology and metabolism 2007;92:2917-22.

10. Hong Y, Liu X, Li Z, Zhang X, Chen M, Luo Z. Real-time ultrasound elastography in the differential diagnosis of benign and malignant thyroid nodules. J Ultrasound Med 2009;28:861-7.

11. Rubaltelli L, Corradin S, Dorigo A, Stabilito M, Tregnaghi A, Borsato S, et al. Differential diagnosis of benign and malignant thyroid nodules at elastosonography. Ultraschall in der Medizin 2009;30:175-9.

12. Kagoya R, Monobe H, Tojima H. Utility of elastography for differential diagnosis of benign and malignant thyroid nodules. Otolaryngol Head Neck Surg 2010;143:230-4.

13. Shuzhen C. Comparison analysis between conventional ultrasonography and ultrasound elastography of thyroid nodules. European journal of radiology 2012;81:1806-11.

14. Trimboli P, Guglielmi R, Monti S, Misischi I, Graziano F, Nasrollah N, et al. Ultrasound sensitivity for thyroid malignancy is increased by real-time elastography: a prospective multicenter study. The Journal of clinical endocrinology and metabolism 2012;97:4524-30.

15. Moon HJ, Sung JM, Kim EK, Yoon JH, Youk JH, Kwak JY. Diagnostic performance of gray-scale US and elastography in solid thyroid nodules. Radiology 2012;262:1002-13.

16. Azizi G, Keller J, Lewis M, Puett D, Rivenbark K, Malchoff C. Performance of elastography for the evaluation of thyroid nodules: a prospective study. Thyroid : official journal of the American Thyroid Association 2013;23:734-40.

17. Unluturk U, Erdogan MF, Demir O, Gullu S, Baskal N. Ultrasound elastography is not superior to grayscale ultrasound in predicting malignancy in thyroid nodules. Thyroid : official journal of the American Thyroid Association 2012;22:1031-8.

18. Mehrotra P, McQueen A, Kolla S, Johnson SJ, Richardson DL. Does elastography reduce the need for thyroid FNAs? Clinical endocrinology 2013;78:942-9.

19. Shweel M, Mansour E. Diagnostic performance of combined elastosonography scoring and high-resolution ultrasonography for the differentiation of benign and malignant thyroid nodules. European journal of radiology 2013;82:995-1001.

20. Kim I, Kim EK, Yoon JH, Han KH, Son EJ, Moon HJ, et al. Diagnostic role of conventional ultrasonography and shearwave elastography in asymptomatic patients with diffuse thyroid disease: initial experience with 57 patients. Yonsei medical journal 2014;55:247-53.

21. Ko SY, Kim EK, Sung JM, Moon HJ, Kwak JY. Diagnostic Performance of Ultrasound and Ultrasound Elastography with Respect to Physician Experience. Ultrasound in medicine & biology 2013.

22. Park SH, Kim SJ, Kim EK, Kim MJ, Son EJ, Kwak JY. Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules. AJR. American journal of roentgenology 2009;193:W416-23.

23. Calvete AC, Rodriguez JM, de Dios Berna-Mestre J, Rios A, Abellan-Rivero D, Reus M. Interobserver agreement for thyroid elastography: value of the quality factor. J Ultrasound Med 2013;32:495-504.

24. Ragazzoni F, Deandrea M, Mormile A, Ramunni MJ, Garino F, Magliona G, et al. High diagnostic accuracy and interobserver reliability of real-time elastography in the evaluation of thyroid nodules.

Ultrasound in medicine & biology 2012;38:1154-62.

25. Merino S, Arrazola J, Cardenas A, Mendoza M, De Miguel P, Fernandez C, et al. Utility and interobserver agreement of ultrasound elastography in the detection of malignant thyroid nodules in clinical care. AJNR Am J Neuroradiol 2011;32:2142-8.

26. Zhang YF, Xu HX, He Y, Liu C, Guo LH, Liu LN, et al. Virtual touch tissue quantification of acoustic radiation force impulse: a new ultrasound elastic imaging in the diagnosis of thyroid nodules. PloS one 2012;7:e49094.

27. Veyrieres JB, Albarel F, Lombard JV, Berbis J, Sebag F, Oliver C, et al. A threshold value in Shear Wave elastography to rule out malignant thyroid nodules: a reality? European journal of radiology 2012;81:3965-72.

28. Bhatia K, Tong CS, Cho CC, Yuen EH, Lee J, Ahuja AT. Reliability of Shear Wave Ultrasound Elastography for Neck Lesions Identified in Routine Clinical Practice. Ultraschall in der Medizin 2012;33:463-8.

29. Cibas ES, Ali SZ. The Bethesda System For Reporting Thyroid Cytopathology. American journal of clinical pathology 2009;132:658-65.

30. Gwet K. Computing inter-rater reliability with the SAS system. Stat Methods Inter-rater Reliability Assess 2002;3:1-16.

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

32. Kim KA, Kim MJ, Jeon HM, Kim KS, Choi JS, Ahn SH, et al. Prediction of microvascular invasion of hepatocellular carcinoma: usefulness of peritumoral hypointensity seen on gadoxetate disodium-enhanced hepatobiliary phase images. Journal of magnetic resonance imaging : JMRI 2012;35:629-34.

33. 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;247:762-70.

34. Tae HJ, Lim DJ, Baek KH, Park WC, Lee YS, Choi JE, et al. Diagnostic value of ultrasonography to distinguish between benign and malignant lesions in the management of thyroid nodules. Thyroid : official journal of the American Thyroid Association 2007;17:461-6.

35. Koike E, Noguchi S, Yamashita H, Murakami T, Ohshima A, Kawamoto H. Ultrasonographic characteristics of thyroid nodules: prediction of malignancy. Archives of surgery 2001;136:334-7.

36. Bhatia KS, Rasalkar DP, Lee YP, Wong KT, King AD, Yuen HY, et al. Cystic change in thyroid nodules: a confounding factor for real-time qualitative thyroid ultrasound elastography. Clin Radiol 2011;66:799-807.

37. Rago T, Santini F, Scutari M, Pinchera A, Vitti P. Elastography: new developments in ultrasound for predicting malignancy in thyroid nodules. J Clin Endocrinol Metab 2007;92:2917-22.

ABSTRACT(IN KOREAN)

탄성 초음파를 이용한 갑상선 결절의 진단 : 관찰자간 신뢰도 평가 및 진단적 유용성

<지도교수 곽 진 영>

연세대학교 대학원 의학과

고 지 은

목적 : 갑상선 결절의 진단에서 탄성초음파의 관찰자간 신뢰도를 평가하고 탄성초음파가 회색조초음파와 추가적으로 사용되었을 때의 추가적 진단적 유용성을 알아보고자 한다. 대상 및 방법 : 총 443개의 병리학적으로 양성 혹은 악성 결절의 회색조초음파 사진과 탄성초음파 동영상 파일을 3개월의 간격을 두고 세명의 다른 기관의 전문의가 평가하였다. 회색조초음파와 Asteria 혹은 Rago 기준을 이용한 탄성초음파의 관찰자간 신뢰도를 평가하였다. 또한 각 관찰자에서 회색조초음파만 사용한 경우와 회색조초음파에 탄성초음파를 추가적으로 사용하였을 때 악성 결절의 진단률을 비교하였다. 결과 : 관찰자간 신뢰도는 회색조초음파와 Asteria 기준을 이용한 탄성초음파 간에 통계학적으로 유의한 차이가 없었다. 하지만 Rago 기준을 이용한 탄성초음파는 가장 낮은 관찰자간 신뢰도를 보였다 (P<0.043). 모든 세명의 관찰자에서 회색조초음파만 이용하였을 때보다 (70.4-80.8%) 탄성초

음파를 추가적으로 사용하였을 때 (81.3-88.3%, 75.4-85.4%) 민감도는 증가하였다. 특이도는 회색조초음파만 이용하였을 때에 (69.0-82.8%) 비해서 탄성초음파를 추가로 사용하였을 때 (51.7-59.1, 59.1-73.9%) 감소하였다. 결론: 탄성초음파는 회색조초음파와 비교하였을 때 비슷한 관찰자간 신뢰도를 보였다. 하지만 회색조초음파만 사용하였을 때 보다 추가적으로 회색조초음파에 탄성초음파를 사용하였을 때 부가적인 진단적 가치는 제한적이다.

핵심되는 말 : 탄성초음파, 갑상선 결절, 관찰자간 신뢰도, 초음파