Ultrasound in Med. & Biol., Vol. ■, No. ■, pp. 1–9, 2015

0301-5629/$ - see front matter

http://dx.doi.org/10.1016/j.ultrasmedbio.2015.08.005

● *Original Contribution*

VARIABILITY IN INTERPRETATION OF ULTRASOUND ELASTOGRAPHY AND GRAY-SCALE ULTRASOUND IN ASSESSING THYROID NODULES

Jieun Koh,* Hee Jung Moon,* Jeong Seon Park,† Soo Jin Kim,‡ Ha Yan Kim,§ Eun-Kyung Kim,* and Jin Young Kwak*

*Department of Radiology, Research Institute of Radiological Science, College of Medicine, Yonsei University, Seoul, Korea; †Department of Radiology, College of Medicine, Hanyang University, Seoul, Korea; ‡Department of Radiology, College of Medicine, Chung-Ang University, Seoul, Korea; and §Biostatistics Collaboration Unit, College of Medicine, Yonsei University, Seoul, Korea

(*Received* 17 *July* 2014; *revised* 5 *August* 2015; *in final form* 7 *August* 2015)

Abstract—The aim of this study was to validate inter-observer variability for strain ultrasound elastography (USE) and to compare the diagnostic performance of a combination of gray-scale ultrasound (US) and USE with that of gray-scale US. Three observers from different institutions evaluated gray-scale US images and USE video files of 443 cytopathologically proven benign or malignant thyroid nodules over a 3-mo period. Inter-observer variability did not statistically differ between USE using the Asteria criteria and gray-scale US; however, USE using the Rago criteria had the lowest inter-observer agreement ($p < 0.043$). For all three observers, sensitivity was increased by adding USE to gray-scale US (81.3%–88.3%, 75.4%–85.4%) compared with gray-scale US (70.4%–80.8%). Specificity was decreased by adding USE to gray-scale US (51.7%–59.1%, 59.1%–73.9%) compared with gray-scale US (69.0%–82.8%). USE and gray-scale US had comparable inter-observer variability. However, on addition of USE to gray-scale US, the additional diagnostic yield was limited compared with that of gray-scale US alone. (E-mail: docjin@yuhs.ac) © **2015 World Federation for Ultrasound in Medicine & Biology.**

Key Words: **Elastography, Thyroid nodule, Inter-observer variability, Ultrasound, Diagnosis.**

INTRODUCTION

Gray-scale ultrasound (US) is the most sensitive test currently available for detecting thyroid lesions; however, differentiation of benign and malignant nodules is not highly accurate with gray-scale ultrasound (Takashima et al. 1995), and thus, its diagnostic value varies considerably from study to study (Fish et al. 2008; Frates et al. 2006; Kim et al. 2002; Kovacevic and Skurla 2007; Lim et al. 2012). Ultrasound elastography (USE) enables the assessment of tissue consistency by differentiating stiff nodules from soft nodules, and it supplements the diagnostic limitations of gray-scale US (Asteria et al. 2008; Azizi et al. 2013; Bamber et al. 2013; Cantisani et al. 2014; Hong et al. 2009; Kagoya et al. 2010; Kim et al. 2014; Mehrotra et al. 2013; Moon et al. 2012; Rago et al. 2007; Rubaltelli et al. 2009; Shuzhen 2012; Shweel and Mansour 2013; Trimboli et al. 2012; Unluturk et al. 2012; Yoon et al. 2014). Previous studies suggested that with respect to diagnostic performance, USE is better or comparable to gray-scale US when differentiating benign from malignant thyroid nodules (Asteria et al. 2008; Azizi et al. 2013; Cantisani et al. 2014; Hong et al. 2009; Rago et al. 2007; Shuzhen 2012; Trimboli et al. 2012). Diagnostic performance was also improved with the combination of gray-scale US and USE (Shweel and Mansour 2013). Contrary to these positive results, several studies have failed to prove the superiority of USE to gray-scale US (Kagoya et al. 2010; Ko et al. 2014; Moon et al. 2012; Unluturk et al. 2012). Moreover, the combination of USE and gray-scale US was found to be inferior to gray-scale US in certain cases (Moon et al. 2012).

In addition to variable diagnostic performance, another technical issue with USE is limited inter-observer agreement (Kwak and Kim 2014), a problem first reported by Park et al. (2009) for strain USE. However, that study did not use subjective methods for monitoring compression. Other consecutive studies reported

Address correspondence to: Jin Young Kwak, Department of Radiology, Research Institute of Radiological Science, College of Medicine, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, 120-752 Seoul, Korea. E-mail: docjin@yuhs.ac

increased inter-observer agreement using subjective methods for monitoring compression on strain USE (Calvete et al. 2013; Merino et al. 2011; Ragazzoni et al. 2012). Shear wave USE has been reported to have fair to excellent reproducibility in neck lesions, including thyroid nodules, with higher inter-observer agreements compared with strain USE (Bhatia et al. 2012; Veyrieres et al. 2012; Zhang et al. 2012). To date, studies evaluating the inter-observer variability of strain USE have been limited by small sample size and have been performed only by observers from the same institution (Calvete et al. 2013; Merino et al. 2011; Park et al. 2009; Ragazzoni et al. 2012). Therefore, we focused on validating the inter-observer agreement for strain USE as well as gray-scale US by having three radiologists from different institutions compare the diagnostic performance of gray-scale US and a combination of gray-scale US and USE in a relatively large number of thyroid nodules.

METHODS

Patients

The institutional review board approved this retrospective study and required neither patient approval nor informed consent for our review of patient images and records. From November 2011 to January 2012, 583 nodules in 465 consecutive patients underwent fine-needle aspiration (FNA) or staging US with strain USE. We excluded nodules measuring <5 mm or ≥ 30 mm ($n = 65$) and nodules for which cytology results were classified as suspicious malignant ($n = 17$), atypia ($n = 32$), follicular neoplasm ($n = 2$) and non-diagnostic results ($n = 10$) with no further surgical intervention. Among 457 nodules, 194 were pathologically confirmed by surgery, and 263 were cytologically proved to be benign or malignant with no further surgical intervention. Among them, 14 nodules were excluded because of the poor quality of USE video files or gray-scale US images. Finally, 443 nodules in 426 patients were included in this study; among these 426 patients, 17 patients had two nodules. Mean age was 47 ± 12 y; 347 patients were female, and 79 were male. The mean size of the nodules was 11 ± 5.6 mm; 212 nodules were ≤ 10 mm, and the remainder were >10 mm.

Gray-scale US and USE

Gray-scale US was performed initially with a 6- to 14-MHz linear array transducer (EUB-7500, Hitachi Medical, Tokyo, Japan) by seven radiologists with 1 to 15 y of experience in thyroid imaging. Transverse and longitudinal images of thyroid nodules were captured and stored for subsequent image interpretation. After 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 2 mo of experience with the machine and had performed USE on 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 a pressure indicator (range: 2–3) were monitored for optimal image acquisition (Moon et al. 2012). In the split-screen mode, gray-scale US images were displayed on the right, and USE images superimposed on the corresponding gray-scale US images were displayed on the left. USE images were displayed with 256 specific colors for each pixel from a color spectrum of red to blue. The softest component was displayed in red and the stiffest component in blue (Moon et al. 2012). USE images were obtained as video files with more than 5 s of continuous length.

Image interpretation

Stored gray-scale images and USE video files were reviewed by one radiology resident (J.E.K.). Appropriate transverse and longitudinal views of each nodule were manually captured after review of gray-scale US images on PACS (picture archiving and communication system). All clinical data were removed from images. Images of each nodule were assigned random numbers and ordered. Video files less than 5 s long were excluded during the USE video file review. USE video files of each nodule were also assigned random numbers different from those of the gray-scale US images.

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

On the gray-scale US image interpretation report, five features of the thyroid nodules were recorded. The internal composition of nodules was recorded as solid, $<50\%$ cystic, $\geq 50\%$ cystic and a cyst. Echogenicity of