Original paper

Cite as: Dobruch-Sobczak K, Migda B, Krauze A, Mlosek K, Słapa RZ, Wareluk P, Bakuła-Zalewska E, Adamczewski Z, Lewiński A, Jakubowski W, Dedecjus M: Prospective analysis of inter-observer and intra-observer variability in multi ultrasound descriptor assessment of thyroid nodules. J Ultrason 2019; 19: 198–206. doi: 10.15557/JoU.2019.0030

Submitted: 20.01.2019
Accepted: 29.03.2019
Published: 30.09.2019

Prospective analysis of inter-observer and intra-observer variability in multi ultrasound descriptor assessment of thyroid nodules

Katarzyna Dobruch-Sobczak[1,2], Bartosz Migda[3], Agnieszka Krauze[3], Krzysztof Mlosek[3], Rafał Z. Słapa[3], Paweł Wareluk[3], Elwira Bakuła-Zalewska[4], Zbigniew Adamczewski[5,6], Andrzej Lewiński[5,6], Wiesław Jakubowski[3], Marek Dedecjus[7]

[1] *Department of Radiology II, The Maria Skłodowska Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland*
[2] *Department of Ultrasound, Institute of Fundamental Technological Research, PAS, Warsaw, Poland*
[3] *Diagnostic Imaging Department, Medical University of Warsaw, 2nd Faculty of Medicine with the English Division and the Physiotherapy Division, Warsaw, Poland*
[4] *Department of Pathology The Maria Skłodowska Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland*
[5] *Department of Endocrinology and Metabolic Diseases, Medical University of Lodz, Poland*
[6] *Polish Mother's Memorial Hospital-Research Institute, Poland*
[7] *Department of Nuclear Medicine and Endocrine Oncology, The Maria Skłodowska Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland*

Correspondence: Bartosz Migda, Diagnostic Imaging Department, Medical University of Warsaw, 2nd Faculty of Medicine with the English Division and the Physiotherapy Division, Warsaw, Poland; tel. +48 22 326 58 10, fax:+48 22 326 54 79, e-mail: bartoszmigda@gmail.com

DOI: 10.15557/JoU.2019.0030

Keywords
inter-observer variability, intra-observer variability, thyroid nodule, ultrasound, sonoelastography

Abstract

Aim: The aim of this study was to evaluate the inter- and intra-observer variability and accuracy of ultrasound assessment of thyroid nodules using a descriptive lexicon. **Materials and methods:** A prospective study was performed on complete ultrasound examinations, including sonoelastography and color Doppler ultrasound of 18 patients with 20 thyroid nodules. A total of 20 records of thyroid nodules from these techniques were duplicated, numbered, and randomly arranged. Five radiologists assessed the recordings independently. Cohen Kappa and Fleiss Kappa statistics were used to determine the degree of intra- and inter-observer agreement. **Results:** Mean accuracy rates for all radiologists, for all ultrasound features, ranged from 82.7 to 87.8%. For B-mode and strain elastography, accuracies ranged from 65.0 to 100% and 47.4 to 86.8%, respectively. Concerning intra-observer variability, three radiologists demonstrated almost perfect agreement (the κ-value ranged from 0.81 to 0.86), and a substantial agreement was noted for the two remaining radiologists. The κ-values for inter-observer agreement ranged from 0.61 for macrocalcifications (substantial agreement) to 0.33 for Asteria four-point elastography scale criteria (fair agreement). **Conclusions:** The results suggest relatively good inter-observer and excellent intra-observer agreement in the assessment of thyroid nodules using ultrasound, and fair agreement in the case of strain elastography.

Introduction

The worldwide incidence of thyroid cancer is steadily increasing[1]. According to the American Thyroid Association (ATA), thyroid nodules are a common clinical problem, with nearly 68% of all examined adult patients diagnosed with these lesions. In 7–15% of these cases, the nodules were found to be carcinomas[2]. In Poland, 3,529 new cases of thyroid cancer were diagnosed in 2015. The annual incidence rate has increased from 3.8 per 100,000 in 2000 to 9.2 per 100,000 in 2015[3,4].

Although thyroid nodules are a common occurrence, it is usually difficult to detect them without imaging techniques (only 4–7% can be palpated)[5]. Thus, ultrasound (US) examinations play an important role in detection. US is a non-invasive, cost-effective, and widely available technique used to discern specific features of nodules and guide fine-needle aspiration biopsy (FNAB)[5].

Published studies, including the ATA Management Guidelines, have demonstrated that hypoechogenicity, irregular margins, microcalcifications, and a taller-than-wide shape are the B-mode features with the highest level of specificity for the detection of malignant thyroid nodules[6–8]. However, none of these features, taken individually, are exclusive to malignant lesions, and benign nodules with a single abnormal feature are relatively common[2,9–11].

Thus, new, non-invasive imaging methods capable of supporting the differentiation of thyroid lesions are being developed. Recently, sonoelastography has become an increasingly used technique[12,13].

Currently, two main types of elastography are available: shear wave elastography (SWE) and strain elastography (SE). Some authors have suggested that SWE is superior to SE in thyroid nodule stratification because of its operator independency, but recent meta-analyses have surprisingly shown that SE has better diagnostic accuracy than SWE[14,15] In addition, Dighe *et al*. described SWE artifacts and their impact on final results[16]. In this paper, the authors suggested that almost 70% of SWE scans have artifacts. Over 18% of scans were unsuitable for final assessment and over 43% of artifacts from unsuitable SWE evaluation were operator dependent[16]. It is known that SE also has limitations. Results are highly dependent on the presence of calcifications (macro- and rim calcifications) in the tumor, as well as location (deep-lying lesions) and tumor type (papillary thyroid cancers – PTC are often more suspicious than follicular thyroid cancers – FTC). Assessments of thyroid nodule malignancy carried out with strain elastography and US depend on examiner's experience level and are characterized by substantial inter-observer variation[5,17–20], but they are useful in monitoring patients who have undergone FNAB[21–23].

Few studies analyzed inter- and intra-observer variability in US diagnosis and even fewer evaluated variability in the case of sonoelastography[5,17–20]. Therefore, we investigated these two variabilities in thyroid nodule evaluations carried out with US and sonoelastography.

Materials and methods

Patients

In this prospective study, patients first gave informed consent to participate in the research. Then they underwent US examination of the thyroid, followed by US-guided biopsy or surgical procedures. The study protocol was approved by the institutional review board of the Maria Skłodowska-Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland. From February to November 2017, 92 consecutive patients (22 men, 70 women) with a total of 149 thyroid nodules were included and examined in the Department of Oncological Endocrinology and Nuclear Medicine, Maria Skłodowska-Curie Memorial Cancer Centre and Institute of Oncology, Warsaw, Poland. Of these, 18 (4 men, 14 women) patients aged 21–78 years, with a total of 20 thyroid nodules, were randomly selected for the study. The nodules included eight malignant and 12 benign diagnoses. All malignant lesions and eight benign nodules were confirmed by postoperative histopathology. The remaining four benign nodules with CII in cytology were excluded from surgery because it was unethical to operate on patients without any indications. In this group, we performed US follow-up at 6 month intervals (Fig. 1).

The inclusion criteria were the presence of a thyroid nodule that underwent US-guided FNAB (according to the Guidelines of Polish National Societies[23], prepared on the initiative of the Polish Group for Endocrine Tumours and the ATA) and was operated or was under active observation including repeated FNAB. The following criteria excluded nodules from further analysis: pure cystic lesions, eggshell calcifications, or non-diagnostic cytology results. The researchers were blinded to the cytological and/or histological results.

Histology

Fourteen patients underwent thyroidectomy and FNAB while four underwent FNAB only. Histologic and cytological findings were used as study endpoints. For patients with benign FNAB results, a US examination was performed within six months. FNABs were performed with 22- to 24-gauge needles, and aspirates were fixed in 75% ethanol and stained with hematoxylin and eosin (H&E). Lesions were assigned to the Bethesda I–VI category[24] based on FNAB findings. FNAB was repeated for nodules classified as CI (non-diagnostic specimen for example cystic fluid only, acellular specimen), CIII (AUS/FLUS Atypia of Undetermined Significance/Follicular Lesion of Undetermined Significance), and small C IV nodule

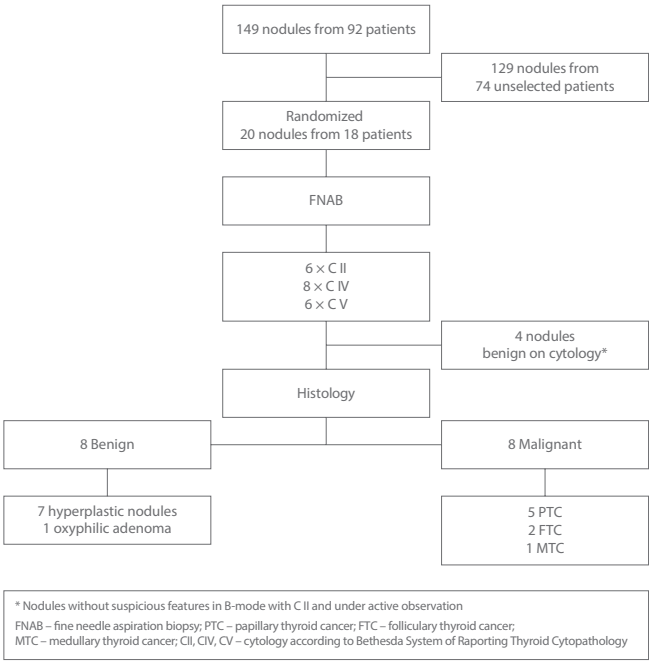

Fig. 1. *Chart flow*

(<1 cm) (Suspicion of Follicular Lesion in small nodules under 1 cm). If possible, a specific histotype was suggested. Cytological results (CV and CVI) were verified by an independent, second pathologist. Surgical specimens were immediately fixed in 10% buffered formalin. Representative sections from these specimens were processed and routinely stained with H&E for histopathologic (microscopic) examinations.

Conventional B-mode and US examinations

Five radiologists (one from oncology centre and four from clinical centre), with experience in thyroid B-mode and CDUS imaging ranging from six to 22 years and experience in US elastography from one to seven years, performed and assessed the examinations. Before the study began, the radiologists were trained in our lexicon: composition; echo pattern in comparison to thyroid parenchyma (Echo-Pa); dominating echo pattern in comparison to thyroid parenchyma – in the case of mixed echogenicity dominating component (Echo-Pb); echo pattern in comparison to muscles (Echo-M); margins; the 'halo' phenomenon; extrathyroidal extension (the observers were asked to determine whether the extrathyroidal extension modeled the shape of the thyroid and its capsule or extended beyond it); macrocalcification; microcalcification; elasticity score (according to Asteria scale), (all features included in Table 1). All examinations were performed with the same protocol, described below.

The US probe was gently placed on the thyroid in a transverse and longitudinal orientation while the patient was in the supine position. The thyroid gland was scanned from superior to inferior in transverse cross sections and from the outer to the inner margin in longitudinal cross sections. The anteroposterior, transverse, and longitudinal measurements of the gland and nodules were taken on frozen images during examination and archived as well. Other B-mode features regarding the lexicon as well as CDUS and SE were assessed retrospectively on archived AVI films and frozen images. CDUS was performed in all cases with the same scale settings (maximal velocity of 2.5 cm/s). The gain of CDUS was adjusted to each patient individually, achieving the appropriate highest sensitivity without blooming artifacts. In the case of SE, since nodules become stiffer during compression, all radiologists avoided pressing the neck with the probe during examinations to minimize false-positive findings. Grey-scale conventional US with CDUS and SE were performed using an iU22 US machine (Philips Medical Systems, Bothell, WA) equipped with a 5–12 MHz linear array transducer. Sonoelastography was assessed qualitatively using Asteria four-point scale criteria (Tab. 1)[25]. The following lesion features were assessed in US and SE examinations (Tab. 1). We excluded shape (taller than wide parameter) of the nodule because the assessment of this feature is more objective as it is done by comparing nodule measurements (height and width, in this research done prospectively). In the case of this research, assessment of inter- and intra-observer agreement including this parameter could have overestimated the final results.

Evaluation of nodules with US by independent observers

From 149 examined thyroid nodules, records of 20 nodules were drawn out. For this purpose, we used MS Excel. The 20 original US records from B-mode, CDUS, and SE were duplicated. The resulting 40 records were numbered and arranged randomly in a final file. All researchers received the same set of files for evaluation. Then, five radiologists evaluated records (AVI loops and JPG images) containing transversal and longitudinal B-mode cross sections of the thyroid lobes. Next, CDUS and SE records (AVI loops and JPG images) of these nodules were assessed.

Statistical analysis

The scoring results for all five observers were calculated using Statistical Software Package (Dell Inc. (2016)), Dell Statistica (data analysis software system, v. 13. software.dell.com). An overall kappa value (κ-value) was estimated for multiple observers. Cohen's kappa coefficient was used to determine the degree of intra-observer agreement, after correcting for agreement expected by chance, between duplicated records of the same patient. For inter-observer agreement, Fleiss kappa statistic was used. Again, correction was made for agreement expected by chance[26].

The kappa values were interpreted according to Landis and Koch[27], i.e., $\kappa < 0.00$ corresponds to poor agreement,

Tab. 1. *List of ultrasound descriptors for thyroid nodules*

US features	Abbreviations	Characteristics
Composition	Composition	Cystic Spongiform Mixed cystic (≥50% cystic volume) Mixed solid (≥50% solid volume) Almost completely solid Solid Cannot determine
Echo pattern (in comparison to thyroid parenchyma)	Echo-Pa	Isoechoic Mixed Hypoechoic
Dominating echo pattern (in comparison to thyroid parenchyma)	Echo-Pb	Hyperechoic Isoechoic Hypoechoic
Echo pattern (in comparison to muscles)	Echo-M	Hyperechoic Isoechoic Hypoechoic
Margins	Margins	Well-defined Ill-defined
"Halo" pattern	"Halo"	Complete Partial Absent
Extrathyroidal extension	Capsule	Models thyroid shape and capsule Invasion beyond the thyroid capsule Absent
Macrocalcifications (>1 mm)	Macro	Present Absent
Microcalcifications (≤1 mm)	Micro	Present Absent
Vascularity	Vascularity	Peripheral Central Combined (central and peripheral) Absent
Elasticity score (Asteria Scale)	Asteria Scale	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
Remaining thyroid parenchyma	Parenchyma	Homogeneous Heterogeneous
Autoimmune thyroiditis	AT	Present Absent
Parenchyma vascularity	Parenchyma vascularity	Normal Decreased Increased

κ = 0.00–0.20 to slight agreement, κ = 0.21–0.40 to fair agreement, κ = 0.41–0.60 to moderate agreement, κ = 0.61–0.80 to substantial agreement, and κ = 0.81–1.00 to almost perfect agreement.

Finally, the accuracies of all researchers were assessed and compared, and the mode was determined for every descriptor in this set of data. This value was assumed to be the correct one for a given descriptor. Researchers who agreed with this value were given an accuracy value of 1; the rest were given an accuracy value of 0. Next, the total accuracy score for every researcher was calculated independently for every descriptor.

Results

Our randomly selected group consisted of 20 nodules in 18 patients (12 nodules were benign, 8 were malignant). The maximum length of the tumors ranged from 6 to 46 mm (mean length 9.7 ± 5.6 mm). Five of them were PTC (papillary thyroid cancer), two were FTC (follicular thyroid cancer) and one was MTC (medullary thyroid cancer). In the benign group, eight were verified by histology and most (7/8) of them were hyperplastic nodules (Fig. 1).

The results of accuracy assessment of the five radiologists are presented in Table 2. The mean accuracy rates for all radiologists for all features ranged from 82.7 to 87.8%. All radiologists achieved accuracy rates ranging from 65.0 to 100% for B-mode examination, and from 47.4 to 86.8% for SE. The highest level of accuracy among all observers was noted when the following features were analyzed: macrocalcifications (from 90.0 to 100%), microcalcifications (from 85.0 to 100%), and evaluation of echo pattern in comparison to strap muscles (from 87.5 to 95. 0%). The intra- and inter-observer variabilities for US, CDUS, and SE features of thyroid nodules are presented in Table 3.

Concerning intra-observer variability, almost perfect agreement was noted for three observers: the second, third, and fourth observers achieved mean κ-values of 0.82, 0.86, and 0.81, respectively. However, substantial agreement in mean κ-values was also noted for the first and fifth observer. Inter-observer agreement, demonstrated by κ-values, ranged from 0.61 for macrocalcifications (substantial agreement) to 0.33 for Asteria criteria (fair agreement).

Inter-observer variability for the majority of US features showed moderate agreement in the estimation of composition (κ = 0.55), echo pattern (Echo-Pa, Echo-Pb, Echo-M) (κ ranging from 0.48 to 0.50), capsule assessment (κ = 0.40) (Fig. 3A), and microcalcifications (κ = 0.57) (Fig. 2). When assessing vascularity, overall agreement was fair (κ = 0.34). The mean inter-observer agreement for all US and SE features was 0.42, corresponding to moderate agreement (Fig. 3B).

Tab. 2. *Assessment of intra-observer accuracy*

Description	Accuracy (%)				
	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5
Composition	92.5	95.0	77.5	72.5	90.0
Echo-Pa	80.0	80.0	90.0	90.0	82.5
Echo-Pb	92.5	82.5	87.5	95.0	87.5
Echo-M	87.5	95.0	90.0	92.5	97.5
Margins	82.5	87.5	72.5	82.5	92.5
"Halo"	85.0	87.5	82.5	65.0	87.5
Capsule	87.5	85.0	80.0	92.5	92.5
Macro	100.0	92.5	97.5	90.0	97.5
Micro	100.0	92.5	85.0	87.5	95.0
Vascularity	65.0	95.0	85.0	97.5	75.0
Asteria Scale	78.9	86.8	68.4	47.4	73.7
Parenchyma	82.5	82.5	80.0	87.5	65.0
AT	77.5	80.0	92.5	97.5	100.0
Parenchyma vascularity	75.0	87.5	70.0	77.5	85.0
Average	84.7	87.8	82.7	83.9	87.2

Discussion

Ultrasonography is a widely accepted imaging technique that accurately detects thyroid nodules and architectural distortion. Over the past decade, significant improvements have been made in US machine technology and high-resolution probes. Therefore, US features specific to thyroid tumors such as lesion stiffness, microcalcification, vascular pattern or margins, can be observed with high accuracy. These US features enable better stratification of malignancy risk and were used to create several Thyroid Imaging Reporting and Data System (TIRADS) classifications, although none were used in clinical practice in Poland[9,28–32]. The primary objective of our study was to evaluate inter- and intra-observer agreement for the selected US and SE features as a first step towards proposing the TIRADS classification.

Calcifications

In our study, the most substantial agreement was obtained when macrocalcifications were evaluated: κ was 0.61 for inter-observer agreement and between 0.64 and 1.0 for intra-observer agreements. For microcalcifications, characterized by stronger associations with tumor malignancy than macrocalcifications, we achieved moderate agreement[33]. Therefore, we assessed them separately in our study. Our results are similar to those reported by Park *et al.*[18], who used the same definition: microcalcification ≤1 mm, macrocalcification >1 mm. In this study, five radiologists, each with one to six years of experience in the assessment of thyroid nodules, received κ-values of 0.40 for macrocalcifications and 0.54 for microcalcifications.

In another study, in which Park *et al.*[20] evaluated thyroid carcinomas only (51 of 52 were PTCs), calcifications were observed in over half of the nodules (depending on observer, ranged between 34 and 42 of 52 nodules; 65.4–80.7%). These authors achieved similar results to those presented here, with κ-values ranging from 0.47 to 0.62 for all types of calcifications. In our study, microcalcifications were found in 7 of 8 thyroid carcinomas, while macrocalcifications were present in 2 of 8 thyroid carcinomas. Moreover, the rate of agreement in the assessment of calcifications in our study was comparable to the results of Choi *et al.*[5] The presence of microcalcifications in the sonographic pattern indicates the need

Tab. 3. *Assessment of the intra- and inter-observer agreements*

Description	Intra-observer agreement										Inter-observer agreement
	Observer 1		Observer 2		Observer 3		Observer 4		Observer 5		
	Agreement (%)	κ-value (SE)	Agreement (%)	κ-value (SE)	Agreement (%)	κ-value (SE)	Agreement (%)	κ-value (SE)	Agreement (%)	κ-value (SE)	κ-value (SE)
Composition	95.0	0.88 (0.12)	100.0	1.00 (0.00)	90.0	0.83 (0.11)	95.0	0.91 (0.09)	90.0	0.75 (0.17)	0.55 (0.04)
Echo-Pa	75.0	0.57 (0.16)	100.0	1.00 (0.00)	90.0	0.83 (0.11)	100.0	1.00 (0.00)	75.0	0.53 (0.17)	0.48 (0.04)
Echo-Pb	85.0	0.69 (0.16)	95.0	0.91 (0.09)	95.0	0.89 (0.11)	100.0	1.00 (0.00)	80.0	0.52 (0.21)	0.50 (0.05)
Echo-M	90.0	0.80 (0.14)	100.0	1.00 (0.00)	100.0	1.00 (0.00)	95.0	0.84 (0.16)	90.0	0.67 (0.18)	0.49 (0.04)
Margins	85.0	0.70 (0.16)	95.0	0.90 (0.10)	95.0	0.88 (0.12)	95.0	0.88 (0.12)	85.0	0.63 (0.20)	0.39 (0.05)
"Halo"	80.0	0.68 (0.13)	85.0	0.76 (0.12)	85.0	0.77 (0.12)	90.0	0.67 (0.19)	75.0	0.62 (0.14)	0.41 (0.04)
Capsule	80.0	0.64 (0.17)	95.0	0.91 (0.09)	85.0	0.72 (0.15)	80.0	0.59 (0.18)	75.0	0.50 (0.19)	0.40 (0.04)
Macro	100.0	1.00 (0.00)	95.0	0.83 (0.17)	95.0	0.64 (0.33)	100.0	1.00 (0.00)	95.0	0.64 (0.33)	0.61 (0.05)
Micro	95.0	0.89 (0.11)	95.0	0.90 (0.10)	90.0	0.77 (0.15)	90.0	0.78 (0.14)	90.0	0.74 (0.17)	0.57 (0.05)
Vascularity	90.0	0.86 (0.10)	90.0	0.74 (0.16)	100.0	1.00 (0.00)	95.0	0.85 (0.13)	95.0	0.87 (0.13)	0.34 (0.03)
Asteria Scale	78.9	0.71 (0.13)	78.9	0.70 (0.13)	94.7	0.92 (0.08)	73.7	0.61 (0.15)	73.7	0.65 (0.13)	0.33 (0.03)
Parenchyma	75.0	0.50 (0.19)	85.0	0.69 (0.16)	100.0	1.00 (0.00)	95.0	0.89 (0.10)	90.0	0.69 (0.20)	0.40 (0.05)
AT	95.0	0.88 (0.12)	80.0	0.47 (0.23)	95.0	*	95.0	0.64 (0.33)	100.0	1.00 (0.00)	0.25 (0.05)
Parenchyma vascularity	80.0	0.64 (0.16)	85.0	0.66 (0.17)	95.0	0.92 (0.08)	85.0	0.71 (0.15)	75.0	0.17 (0.26)	0.18 (0.04)
Average	86.0	0.74	91.4	0.82	93.6	0.86	92.0	0.81	84.9	0.64	0.42

* The data structure did not allow κ-value and SE to be calculated

Fig. 2. *B-mode US image of solid, inhomogeneous hypoechoic lesion with lobular margins and a size of 21 × 17 × 28.5 mm (depth × width × length). During assessment of microcalcifications, there was disagreement between observers, three were negative, and two of were positive for this feature. In histology, the lesion was proved to be follicular thyroid cancer*

for a biopsy, but more importantly, accurate evaluation of the sample[6].

In this study, assessment of the final results revealed that disagreement in terms of microcalcifications appears in nodules that were more normoechogenic or had hyperechogenic components (in the case of mixed echogenicity where microcalcifications were presented in the hyperechogenic component). This could affect the contrast between spot-like <1 mm reflection and surrounding solid parts of the nodule. Unfortunately, this disagreement was found in three malignant lesions, one PTC, one follicular variant of PTC and one FTC (Fig. 2). The follicular variant of PTC and FTC were normoechogenic, which could decrease contrast mentioned above. PTC was hypoechogenic, but the dimension was under 10 mm, which could be another limitation in the evaluation of microcalcifications.

Echogenicity

In order to assess an echogenic nodule, we compared it with the thyroid parenchyma or the strap muscles, or used the dominant echo pattern. Inter-observer analysis of this parameter revealed moderate agreement (κ-values ranging from 0.48 to 0.5). This result may be partially explained by complex echogenicity of thyroid tumors and the background parenchyma. Data analysis revealed that besides complex echogenicity, the structure of the nodule was also important. More disagreement occurred for nodules with a mixed solid-fluid structure. The size of the nodules was also important. There was more disagreement for large nodules

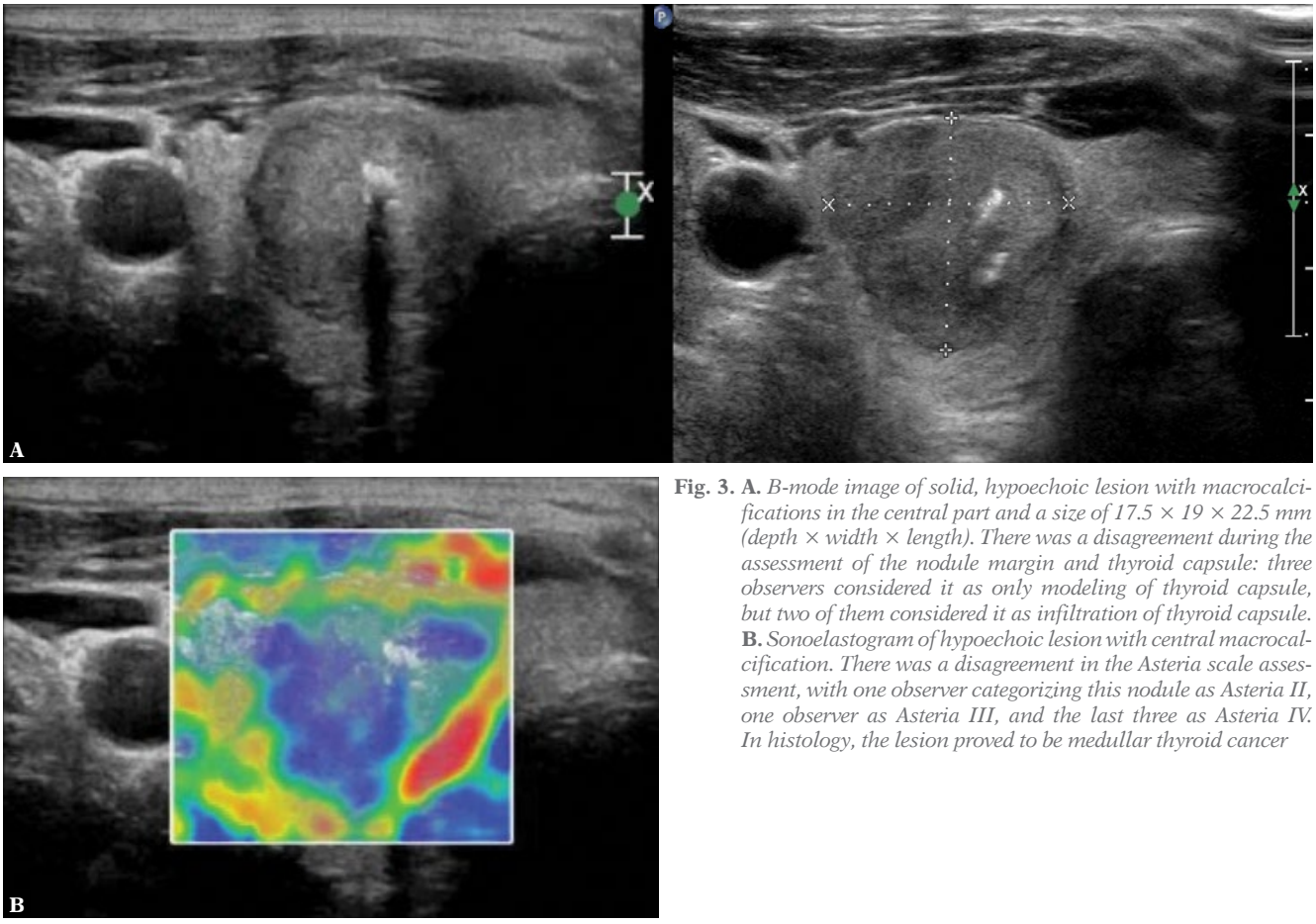

Fig. 3. A. *B-mode image of solid, hypoechoic lesion with macrocalcifications in the central part and a size of 17.5 × 19 × 22.5 mm (depth × width × length). There was a disagreement during the assessment of the nodule margin and thyroid capsule: three observers considered it as only modeling of thyroid capsule, but two of them considered it as infiltration of thyroid capsule.* **B.** *Sonoelastogram of hypoechoic lesion with central macrocalcification. There was a disagreement in the Asteria scale assessment, with one observer categorizing this nodule as Asteria II, one observer as Asteria III, and the last three as Asteria IV. In histology, the lesion proved to be medullar thyroid cancer*

filling the whole lobe than smaller ones in terms of echogenicity in relation to parenchyma. This could be caused by less surrounding parenchyma for comparison. Choi *et al*. found fair agreement when these features were evaluated (κ-values were 0.34 for the first observer, 0.45 for the second), subdividing this category into only four types: hyperechoic, isoechoic, hypoechoic, and marked hypoechoic. On the other hand, we observed almost perfect intra-observer agreement (κ-values ranging from 0.83 to 1), even though we provided a very detailed definition of this feature by assigning it seven possible characteristics (Tab. 1).

Margins

The characteristics of lesion margins are an important feature when evaluating malignancy. When differentiating between benign and malignant thyroid nodules, nodules with circumscribed margins are more likely to be benign. However, this feature has low sensitivity as 33 to 93% of malignant tumors may also have circumscribed margins[34]. It is difficult to identify margins when the surrounding thyroid gland is heterogeneous or borders of the nodules overlapped. The results presented by other researchers demonstrated a high degree of inter-observer variability when nodule margins were assessed[11]. In our study, the margins could be described as either well-circumscribed or not circumscribed (lobular, spiculated, angular, jagged). Evaluation of this feature resulted in the lowest level of inter-observer agreement (κ-value of 0.39) and satisfactory intra-observer agreement (κ-values ranging from 0.65 to 0.90). Choi *et al*. and Park *et al*. used the same categorization of margins and obtained similar results, with κ-values of 0.42 and 0.03–0.29, respectively[5,20]. In our study, most of the disagreement was for nodules positioned tangentially to the thyroid capsule, between the isthmus and lobe or when nodules brought out the capsule. These differences can be caused by uneven thickness of the thyroid capsule in contact with the nodule.

'Halo' phenomenon and capsule invasion

The subsequent features assessed were the 'halo' phenomenon and capsule invasion. Here, observer agreement was moderate, indicating that evaluation of this feature is characterized by limited accuracy. Park *et al*. showed even less favorable results, with only fair agreement (κ-value of 0.32) for determination of capsule invasion[18]. In both the "halo" phenomenon and capsule invasion, we demonstrated disagreement mostly in large nodules that brought out the gland capsule, or nodules that were in contact with capsule (Fig. 3A), or were part of a nodule conglomerate.

Strain elastography

In our study, the determination of lesion stiffness using a 4-grade scale was a difficult task for all observers as the level of agreement was fair. Four radiologists experienced in SE assessment achieved levels of accuracy from 68.4 to 86.8%. One radiologist, with only one year of experience achieved only a 47.4% level of accuracy (Fig. 3B). Inter-observer agreement was fair, with a κ-value of 0.33. This could be caused by different level of experience. In published papers, results vary between research centers[19,20]. Friedrich-Rust *et al*. used the same 4-grade scale for qualitative SE and obtained substantial agreement between three observers (κ = 0.66). In contrast, Park *et al*. obtained only slight agreement between observers for this technique (ranging from κ = 0.08 to κ = 0.22)[20]. However, a meta-analysis of SE from 2010 reported high mean values of sensitivity and specificity for diagnoses of 92% and 90%, respectively[34], testifying to the efficacy of this method, which, in our opinion, could be an important accessory in an experienced hand. In our study all observers assessed the same copies of the original files (AVI videos) along with the B-mode real-time records. Therefore, assessment of the accuracy was greater in comparison to the still images used only by Park CS *et al*. The fair agreement in the case of SE assessment may be associated with the lower experience of one of the researchers in this technique. The results of our study suggest the need for a discussion concerning whether SE, which is still a new and rarely used technique, should be part of the lexicon in further research.

Limitations

Our study had several limitations. We included patients from the Department of Oncological Endocrinology and Nuclear Medicine pre-diagnosed with suspicious nodules or in whom carcinomas were detected. Therefore, the group of patients differed from a general screening population; the proportion of malignant lesions in our group was 45%. This is a general limitation of most studies performed in endocrinology and oncology centers, in which there are generally high percentages of malignant cases. The proposed lexicon was very detailed and despite previous training for all radiologists, some misunderstandings occurred. Our results showed too many US features used for nodule assessment, and further research should re-evaluate them. We used operator-dependent strain elastography, which has some limitations (probe placement in relation to common carotid artery – more noise when probe in transverse section close to CCA (common carotid artery); probe compression; the presence of rim calcifications or multiple macrocalcifications covering the nodule; fluid parts of the nodule). However, in relation to SWE, which is thought to be more independent, recent reports have pointed out that this technique also has operator-dependent artifacts and limitations[16]. Besides that, we used a single US machine and did not compare SE from different US systems. It could be assumed that the use of SE from different companies could cause differences in the final results, but this should be further analyzed in a prospective study. Hence, the US machine software

and hardware should be considered when creating the TIRADS lexicon.

Summary

In this study, five radiologists, each with more than six years of experience in thyroid B-mode imaging, assessed 40 thyroid nodules, with relatively good inter-observer agreement and excellent intra-observer agreement in the assessment of thyroid nodules using US and fair agreement in the case of sonoelastography. The highest disagreement was found for capsule invasion, "halo" phenomenon, and the margins of large nodules especially those filling most of the thyroid lobe and/or found in vicinity of the thyroid capsule. In the case of microcalcifications, the differences appear mostly in normoechogenic nodules or nodules with a hyperechogenic component.

Conclusion

Sonographers must be watchful when assessing margin and capsule invasion in large nodules that are filling a significant part of the lobe or lying near the capsule, as well as when assessing microcalcifications in normoechogenic nodules or with hyperechogenic components.

All results suggest relatively good inter-observer and excellent intra-observer agreement in the assessment of thyroid nodules using US, and fair agreement in the case of sonoelastography.

Conflict of interest

Authors do not report any financial or personal connections with other persons or organizations which might negatively affect the contents of this publication and/or claim authorship rights to this publication.

References

1. Guth S, Theune U, Aberle J, Galach A, Bamberger CM: Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest 2009; 39: 699–706.
2. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE *et al*.: 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016; 26: 1–133.
3. Didkowska J, Wojciechowska U, Olasek P: Nowotwory złośliwe w Polsce w 2015 roku. Centrum Onkologii – Instytut im. Marii Skłodowskiej-Curie, Warszawa 2017.
4. Instytut ZEiPNCO: Krajowy Rejestr Nowotworów. Available from: http://onkologia.org.pl/raporty/#tabela_nowotwor.
5. Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ: Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid 2010; 20: 167–172.
6. Brito JP, Gionfriddo MR, Nofal A, Boehmer KR, Leppin AL, Reading C *et al*.: The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. J Clin Endocrinol Metab 2013; 99: 1253–1263.
7. Campanella P, Ianni F, Rota CA, Corsello SM, Pontecorvi A: Quantification of cancer risk of each clinical and ultrasonographic suspicious feature of thyroid nodules: A systematic review and meta-analysis. Eur J Endocrinol 2014; 170: R203–R211.
8. Remonti LR, Kramer CK, Leitao CB, Pinto LCF, Gross JL: Thyroid ultrasound features and risk of carcinoma: A systematic review and meta-analysis of observational studies. Thyroid 2015; 25: 538–550.
9. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH *et al*.: Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. Radiology 2011; 260: 892–899.
10. Salmaslıoğlu A, Erbil Y, Dural C, İşsever H, Kapran Y, Özarmağan S *et al*.: Predictive value of sonographic features in preoperative evaluation of malignant thyroid nodules in a multinodular goiter. World J Surg 2008; 32: 1948–1954.
11. 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.
12. Szczepanek-Parulska E, Woliński K, Stangierski A, Gurgul E, Biczysko M, Majewski P *et al*.: Comparison of diagnostic value of conventional ultrasonography and shear wave elastography in the prediction of thyroid lesions malignancy. PLoS One 2013; 8: e81532.
13. Szczepanek-Parulska E, Woliński K, Stangierski A, Gurgul E, Ruchała M: Biochemical and ultrasonographic parameters influencing thyroid nodules elasticity. Endocrine 2014; 47: 519–527.
14. Tian W, Hao S, Gao B, Jiang Y, Zhang X, Zhang S *et al*.: Comparing the diagnostic accuracy of RTE and SWE in differentiating malignant thyroid nodules from benign ones: a meta-analysis. Cell Biochem 2016; 39: 2451–2463.
15. Hu X, Liu Y, Qian L: Diagnostic potential of real-time elastography (RTE) and shear wave elastography (SWE) to differentiate benign and malignant thyroid nodules: A systematic review and meta-analysis. Medicine 2017; 96: e8282.
16. Dighe M, Hippe DS, Thiel J: Artifacts in shear wave elastography images of thyroid nodules. Ultrasound Med Biol 2018; 44: 1170–1176.
17. Moon HJ, Yoon JH, Kwak JY, Chung WY, Nam KH, Jeong JJ *et al*.: Positive predictive value and interobserver variability of preoperative staging sonography for thyroid carcinoma. AJR Am J Roentgenol 2011; 197: W324–W330.
18. Park CS, Kim SH, Jung SL, Kang BJ, Kim JY, Choi JJ *et al*.: Observer variability in the sonographic evaluation of thyroid nodules. J Clin Ultrasound 2010; 38: 287–293.
19. Friedrich-Rust M, Meyer G, Dauth N, Berner C, Bogdanou D, Herrmann E *et al*.: Interobserver agreement of Thyroid Imaging Reporting and Data System (TIRADS) and strain elastography for the assessment of thyroid nodules. PLoS One 2013; 8: e77927.
20. 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 Am J Roentgenol 2009; 193: W416–W423.
21. Cosgrove D, Piscaglia F, Bamber J, Bojunga J, Correas JM, Gilja O *et al*.: EFSUMB guidelines and recommendations on the clinical use of ultrasound elastography. Part 2: clinical applications. Ultraschall Med 2013; 34: 238–253.
22. Shiina T, Nightingale KR, Palmeri ML, Hall TJ, Bamber JC, Barr RG *et al*.: WFUMB guidelines and recommendations for clinical use of ultrasound elastography: Part 1: basic principles and terminology. Ultrasound Med Biol 2015; 41: 1126–1147.
23. Jarząb B, Dedecjus M, Słowińska-Klencka D, Lewiński A, Adamczewski Z, Anielski R *et al*.: Guidelines of Polish National Societies Diagnostics and Treatment of Thyroid Carcinoma. 2018 Update. Endokrynol Pol 2018; 69: 34–74.
24. Straccia P, Rossi ED, Bizzarro T, Brunelli C, Cianfrini F, Damiani D *et al*.: A meta-analytic review of the Bethesda System for Reporting Thyroid Cytopathology: has the rate of malignancy in indeterminate lesions been underestimated? Cancer Cytopathol 2015; 123: 713–722.

25. 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 2008; 18: 523–531.
26. Fleiss JL, Levin B, Paik MC: Statistical Methods for Rates and Proportions. John Wiley & Sons 2013.
27. Landis JR, Koch GG: The measurement of observe agreement for categorical data. Biometrics 1977; 33: 159–174.
28. Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A *et al*.: An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. J Clin Endocrinol Metab 2009; 94: 1748–1751.
29. Russ G, Royer B, Bigorgne C, Rouxel A, Bienvenu-Perrard M, Leenhardt L: Prospective evaluation of thyroid imaging reporting and data system on 4550 nodules with and without elastography. Eur J Endocrinol 2013; 168: 649–655.
30. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L: European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. Eur Thyroid J 2017; 6: 225–237.
31. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA *et al*.: ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. J Am Coll Radiol 2017; 14: 587–595.
32. Migda B, Migda M, Migda MS, Slapa RZ: Use of the Kwak Thyroid Image Reporting and Data System (K-TIRADS) in differential diagnosis of thyroid nodules: systematic review and meta-analysis. Eur Radiol 2018; 28: 2380–2388.
33. 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: 1002–1013.
34. Bojunga J, Herrmann E, Meyer G, Weber S, Zeuzem S, Friedrich-Rust M: Real-time elastography for the differentiation of benign and malignant thyroid nodules: a meta-analysis. Thyroid 2010; 20: 1145–1150.