











Submitted:
30.11.2021
Accepted:
27.01.2022
Published:
13.04.2022

Summary of meta-analyses of studies considering lesion size cut-off thresholds for the assessment of eligibility for FNAB and sonoelastography and inter- and intra-observer agreement in estimating the malignant potential of focal lesions of the thyroid gland

Katarzyna Dobruch-Sobczak*¹ , Zbigniew Adamczewski*² ,
Marek Dedejusz³ , Andrzej Lewiński^{4,5} , Bartosz Migda*⁶ ,
Marek Ruchała⁷ , Anna Skowrońska-Szcześniak*⁸ ,
Ewelina Szczepanek-Parulska*⁷ , Klaudia Zajkowska*³ , Agnieszka Żyłka*³ 

¹ Department of Radiology II, Maria Skłodowska-Curie National Research Institute of Oncology, Warsaw, Poland

² Department of Nuclear Medicine, Medical University of Lodz, Lodz, Poland

³ Department of Oncological Endocrinology and Nuclear Medicine, Maria Skłodowska-Curie National Research Institute of Oncology, Warsaw, Poland

⁴ Department of Endocrinology and Metabolic Diseases, Polish Mother's Memorial Hospital – Research Institute, Lodz, Poland

⁵ Department of Endocrinology and Metabolic Diseases, Medical University of Lodz, Lodz, Poland

⁶ Ultrasound Diagnostics Laboratory, Department of Paediatric Radiology, Faculty of Medicine, Medical University of Warsaw, Warsaw, Poland

⁷ Department of Endocrinology, Metabolism and Internal Diseases, Poznan University of Medical Sciences, Poznan, Poland

⁸ Department of Internal Medicine and Endocrinology, Medical University of Warsaw, Warsaw, Poland

* These authors contributed equally to this work

Correspondence: Katarzyna Dobruch-Sobczak; e-mail: katarzyna.dobruch-sobczak@pib-nio.pl

DOI: 10.15557/JoU.2022.0021

Keywords

thyroid cancer;
thyroid ultrasound;
relative strain
sonoelastography;
shear wave
sonoelastography;
internal examiner
compliance

Abstract

Thyroid cancer is a tumour with a steadily increasing incidence. It accounts for 7% to 15% of focal lesions detected by ultrasound, depending on age, gender and other factors affecting its occurrence. Fine-needle aspiration biopsy is an essential method to establish the diagnosis but, in view of its limitations, sonoelastography is seen as a non-invasive technique useful in differentiating the nature of lesions and monitoring them after fine-needle aspiration biopsy. This paper presents a literature review on the role of both sonoelastographic techniques (relative strain sonoelastography, shear wave sonoelastography) to assess the deformability of focal thyroid lesions. Ultrasound examination is a relatively subjective method of thyroid imaging, depending on the skills of the examiner, the experience of the centre, and the quality of equipment used. As a consequence, there are inconsistencies between the results obtained by different examiners (*inter-observer variability*) and by the same examiner (*intra-observer variability*). In this paper, the authors present a review of the literature on inter-observer and intra-observer variability in the assessment of individual features of ultrasound imaging of focal lesions in the thyroid. In addition, the authors report on an analysis of cut-off thresholds for the size of lesions constituting the basis for fine-needle aspiration biopsy eligibility assessment. The need to diagnose carcinomas up to 10 mm in diameter is highlighted, however a more liberal approach is recommended in terms of indications for biopsy in lesions associated with a low risk of malignancy, where, based on consultations with patients, active ultrasound surveillance might even be considered.

Introduction

This paper, which is the second part of a meta-analysis review, presents the role of sonoelastography in the differential diagnosis of focal thyroid lesions as a new technique to assess the deformability of thyroid nodules (TN). Ultrasound examination (US) is a relatively subjective method of thyroid imaging, depending on the skills of the examiner, the experience of the centre and the quality of the equipment used^(1,2). As a consequence, inconsistencies may occur between the results obtained by different examiners (*inter-observer variability*) and by the same examiner (*intra-observer variability*). In this paper, the authors present also a review of the literature on inter-observer and intra-observer variability in the assessment of different features of ultrasound imaging of focal lesions in the thyroid gland, and an analysis of lesion size cut-off thresholds constituting the basis for FNAB eligibility assessment.

Sonoelastography: a literature review

Thyroid cancer is a tumour with a steadily increasing incidence. Of the focal lesions detected by ultrasonography, it accounts for 7% to 15%, depending on age, gender and other factors affecting its occurrence. FNAB is an essential method for establishing the diagnosis, but on account of its limitations (false positive, negative, non-diagnostic results) sonoelastography is seen as a non-invasive technique useful in differentiating the nature of lesions and monitoring them after FNAB^(3,4). Malignant lesions have been shown to deform less than most benign lesions. Relative *strain elastography* (SE), one of the two main types, requires compression of the tissues to be evaluated by means of an imaging head or uses arterial pulse or respiratory movements. As such, it is a technique dependent on the experience of the examiner. On the other hand, *shear wave elastography* (SWE) is a new-generation technique that uses an acoustic pulse force to generate a transverse wave, the velocity of which is measured in the tissue and used for its characterisation. Quantitative measurements of tissue stiffness are expressed in kilopascals (kPa) or metres per second (m/s)⁽⁴⁾. This type of sonoelastography, unlike SE, does not require compression and relies less on the examiner than SE. Sonoelastography is recommended by the European Federation of Societies in Ultrasound and Biology (EFSUM), World Federation for Ultrasound in Medicine and Biology (WFUMB), and Polish scientific societies⁽⁴⁻⁶⁾, even though it is not a feature required to assign a category in the Thyroid Imaging Reporting and Data System (TIRADS) category based on the ACR-TIRADS, EU-TIRADS or K-TIRADS systems. However, the deformability of focal thyroid lesions as an independent feature differentiating the nature of focal lesions has been evaluated in numerous publications and meta-analyses. In the 2018 EFSUMB (European Federation of Societies for Ultrasound in Medicine and Biology) guidelines, it is recommended as a useful technique for differentiating the nature of focal lesions and monitoring

lesions verified as benign. If deformability is reduced, it is a predictor of increased risk of malignancy with a recommendation for biopsy.

For their 2020 meta-analysis⁽⁷⁾, the authors included papers in which studies were performed on three different available SWE ultrasound machines: *SuperSonic shear wave elastography* (2D-SWE; Aix-en Provence, France), *Virtual Touch imaging and quantification* (VTIQ; Siemens Medical Solutions, Mountain View, CA) and *Toshiba shear wave elastography* (T-SWE; Toshiba Medical Systems, Tochigi, Japan). A total of 26 studies from 2010–2017 were included in the meta-analysis (verification of lesions by FNAB/observation or histopathological verification), with a total of 3,806 focal lesions analyzed, of which 2,428 were benign and 1,378 malignant. The 2D-SWE technique dominated (10 publications), followed by four papers on the VTIQ technique and three papers on the T-SWE technique. The results of statistical analysis are summarised in Tab. 1. In conclusion, the authors note that the results obtained using the 2D-SWE technique may be an independent predictor of TNs risk.

In a subsequent meta-analysis⁽⁸⁾ and literature review, the authors assessed the diagnostic value of the 2D-SWE method alone. They analysed a total of 2,851 focal thyroid lesions (1,092 malignant and 1,759 benign) based on 14 papers, six of which were included in the previous meta-analysis. Malignant neoplastic lesions accounted for 38.3% of all lesions. The overall sensitivity and specificity and AUC (*area under curve*) were: 0.66 (95% CI: 0.64–0.69), 0.78 (95% CI: 0.76–0.80), and 0.851, and were similar to those reported in Tab. 1. The authors noted the relatively low diagnostic sensitivity of the technique and the high heterogeneity of results. The range of the cut-off point values between the benign and malignant lesions was extensive. For the mean values, it varied from 18.7 kPa to 56.1 kPa (for benign lesions) and from 31.69 kPa to 174 kPa (for malignant lesions). In another meta-analysis⁽⁹⁾ in which the authors evaluated both elastographic techniques, a total of 2,063 benign lesions and 598 malignant neoplastic lesions, verified by histopathological examination, were assessed. For SE (12 papers), the authors found an overall sensitivity of 0.84 (95% CI, 0.76; 0.90) and specificity of 0.9 (95% CI, 0.85; 0.94),

Tab. 1. Summary of sensitivity and specificity and SROC for individual SWE subtypes

	T-SWE	VTIQ	2D-SWE
Sensitivity, % (95% CI)	0.77 (0.70–0.83)	0.72 (0.67–0.77)	0.63 (0.59–0.66)
Specificity, % (95% CI)	0.76 (0.72–0.81)	0.81 (0.78–0.84)	0.81 (0.79–0.83)
SROC	0.84	0.85	0.88
T-SWE – Toshiba shear wave elastography; VTIQ – Virtual Touch imaging and quantification; 2D-SWE – SuperSonic shear wave elastography; SROC – summary receiver operating characteristic			

which is significantly higher than in the conventional ultrasound technique. For SWE (10 papers, 2 papers were included in the Nattabi HA meta-analysis, 1 in the Filho RHC meta-analysis) a sensitivity of 0.79 (95% CI, 0.73; 0.84) and specificity of 0.87 (95% CI, 0.79; 0.92) were obtained. The AUC for SE was 0.94 (95% CI, 0.91; 0.96), while for SWE it was 0.83 (95% CI, 0.80; 0.86), respectively. The difference was statistically significant ($p < 0.01$). The authors noted the higher SE accuracy and specificity compared to the SWE technique.

In their section on limitations, the authors of the meta-analyses highlight the high percentage of malignant lesions, varied dimensions of the lesions, selected groups of patients referred for surgery or FNAB, and differences in the examination technique used, relating primarily to the use of compression. Moreover, the majority of malignant lesions were papillary carcinomas.

Size cut-off thresholds in the assessment of FNAB eligibility: a literature review

The use of *ultrasound risk stratification systems* (US RSSs) serves to categorise focal thyroid lesions according to their ultrasound pattern. These systems, irrespective of the adopted qualification principle, divide ultrasound evaluated lesions into groups from the lowest to the highest risk of malignancy. The most important aim of using US RSSs is to reveal lesions with the highest risk of malignancy. Depending on the category to which a focal lesion is assigned, the risk ranges from 0 to 90%. It should be emphasised, however, that the US features underlying the qualification to the category of high risk of malignancy – repeatedly discussed in the paper – in practice refer to the US features of papillary carcinoma and, to a lesser extent, of medullary carcinoma. Unfortunately, based on US imaging, the authors are unable to categorise cases of follicular carcinoma into high-risk groups, especially in cases of microinvasive carcinomas. Considering the prevalence of each type of cancer in the population (PTC (*papillary thyroid carcinoma*): ca. 85%, MTC (*medullary thyroid carcinoma*): 3–5%, FTC (*follicular carcinoma*): 2–5%, PDTC (*poorly differentiated thyroid carcinoma*): 6%, ATC (*anaplastic thyroid carcinoma*): 1%))⁽¹⁰⁾, it should be emphasised that even though this is a minor limitation to the widespread use of US RSSs, it must be widely known among examiners performing thyroid US. Thus, the categorisation of a focal lesion as high risk should be considered as an indication for FNAB. Another, no less important, purpose of using US RSSs is the ability to detect lesions with a benign US pattern or low risk of malignancy. This translates, in practice, into a reduction in the number of FNABs performed, an effect that some researchers consider to be a more important benefit of using US RSSs over typing malignant lesions. The key issue in this context, in addition to the definition of individual risk categories, becomes the assignment of appropriate cut-off thresholds for the size of lesions that are the basis for FNAB eligibility assessment.

Assuming that FNAB is performed for all focal lesions, a sensitivity of up to 100% is achieved (all malignancies detectable by FNAB are detected), but the specificity of the examination will be very low, which in practice indicates a very high number of biopsies performed in lesions with a benign US pattern.

Introducing restrictions on biopsy eligibility will always affect all assessed statistical parameters to a greater or lesser extent, and have a practical impact on the number of cancers diagnosed and the proportion of unnecessary biopsies performed. In the 2019 paper by Dobruch-Sobczak *et al.*⁽¹¹⁾, adopting cut-off thresholds for the size of focal lesions eligible for FNAB according to the EU-TIRADS classification resulted in a situation where 35% (81/229) of thyroid cancers would not undergo cytological verification. However, it needs to be strongly emphasised that 33.8% of these cases (72/213) involved lesions classified as category 5 and <1 cm. Ha *et al.* in their study⁽¹²⁾ based on a retrospective analysis of 3,323 thyroid nodules showed that the risk of cancer in lesions <1 cm was almost twice higher compared to lesions measuring >1 cm (62.5% (535 of 856) vs 37.5% (321 of 856); $p < 0.001$). In their study, the authors conducted a simulation study to evaluate changes in diagnostic efficiency and the proportion of unnecessary biopsies depending on the cut-off threshold used for the size of TNs eligible for biopsy. They compared the sensitivity, specificity, accuracy, and percentage of unnecessary biopsies that characterise the ATA 2015 and KTA/KSThR 2016 systems with the diagnostic efficiency of six simulations differing in the multiplicity of biopsy eligibility cut-off thresholds in each risk category. The most spectacular change observed in the study was a decrease in sensitivity of more than 20% (ATA 2015 92.5% vs 67%; KTA/KSThR 2016 93.5% vs 66.4%) after increasing the cut-off threshold from 1.0 to 1.5 cm in the intermediate risk categories. This was unrelated to increasing the cut-off threshold to 2.5 cm in the low/very low risk categories according to ATA 2015 and low risk in benign lesions according to KTA/KSThR 2016, which increased the specificity in ATA 2015 from 34.0% to 47.7% and in the KTA/KSThR of 2016 28.7% to 56.3%, while significantly reducing the percentage of unnecessary biopsies: in the ATA 2015 scale from 55.1% to 43.6% and in the KTA/KSThR 2016 from 59.5% to 36.4%.

These data indicate the need to adopt a different priority depending on the ultrasound risk group. This should translate into striving for maximum sensitivity in the high-risk group, as the greatest number of malignant lesions is detected regardless of the rate of unnecessary biopsies, while in the low-risk group, maximum specificity is sought by reducing the number of cytological examinations performed. The treatment of differentiated thyroid carcinomas has recently changed, in particular in cases of papillary carcinoma at stage T1a, which is no longer an absolute indication for total thyroid resection. The adoption of a cut-off threshold of 10 mm in the high-risk group significantly reduces the chances of a less extensive operation, such as thyroid lobectomy

with isthmus, increasing the risk of recurrence and death⁽¹³⁾. This indicates the need to diagnose cancers up to 10 mm in diameter, while taking a more liberal approach regarding the indications for biopsy in lesions with a low risk of malignancy. In such cases, based on consultations with the patient, management could even be restricted to active ultrasound surveillance.

Inter- and intra-observer agreement in the assessment of individual ultrasound imaging features of focal thyroid lesions

The inter-observer agreement is most commonly expressed as the kappa coefficient (Cohen's kappa, Fleiss' kappa, Randolph's kappa)⁽¹⁴⁾. Less commonly used are Krippendorff alpha⁽¹⁵⁾ and *intraclass correlation coefficient* (ICC)⁽¹⁶⁾. The interpretation of the most commonly used kappa coefficient (kappa values) according to Landis and Koch⁽¹⁷⁾ is shown in Tab. 2.

A number of studies have been published on the inter-observer agreement in the assessment of individual thyroid ultrasound imaging features, which ranges from poor to almost perfect, depending on the feature and study^(18–33). Liu *et al.*⁽³⁴⁾ conducted a meta-analysis of seven studies assessing inter-observer agreement published up to December 2018, including a total of 927 patients^(18–23). They calculated the pooled agreement between examiners in the assessment of individual features in thyroid ultrasound images, with the following results: substantial agreement for structure (0.61; 95% CI: 0.55–0.66) and presence of calcifications (0.71; 95% CI: 0.65–0.77), moderate agreement for echogenicity (0.58; 95% CI: 0.51–0.64), shape (0.53; 95% CI: 0.45–0.62), and the presence of echogenic foci, including punctate echogenic foci/microcalcifications, macrocalcifications, peripheral calcifications, and comet tail artefacts (0.43; 95% CI: 0.32–0.54), and fair agreement for margins (0.40; 95% CI: 0.32–0.48). The inter-observer agreement depended, among other factors, on the professional experience of the examiners⁽³⁴⁾. The majority of studies included in the meta-analysis were single-centre studies^(19–23).

An overview of studies published since January 2019 (not included in the Liu *et al.* meta-analysis) is presented in Tab. 3. The studies presented are difficult to compare in view of differences in study design (retrospective/prospective, single-centre/multi-centre, varying number of examiners, different percentage of nodules verified as benign and malignant by biopsy or histopathology, various features assessed by ultrasound)^(24–32), nevertheless some conclusions can be drawn. As shown in Tab. 3, some features in the ultrasound image were characterised by higher agreement than others. The feature with the highest inter-observer agreement was nodule structure, assessed in most studies as high^(24,27,30,31) or moderate^(25,26,29,31). In contrast, the assessment of the presence of comet tail artefacts was characterised by the lowest level of agreement, rated as slight in the majority of studies^(26,29,32). The degree of inter-observer agreement for margins was not much

Tab. 2. Interpretation of the kappa coefficient values according to Landis and Koch⁽¹⁷⁾

Range of kappa values	Interpretation of the degree of agreement
<0,00	poor
0.00–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	almost perfect

better, rated as slight^(28,30,32), fair^(25,26,29) or moderate^(24,27,31). These findings are consistent with the outcomes of the meta-analysis conducted by Liu *et al.*⁽³⁴⁾, and thus demonstrate a distinctive general trend.

The degree of agreement between observers depends on their professional experience^(2,21,34) and improves after training sessions involving joint viewing of ultrasound images and discussions to reach agreement held among participating examiners^(2,19,31).

Far fewer studies are available to assess the intra-observer agreement for individual thyroid ultrasound features^(22,25,29,33). The reproducibility of the results of repeat examinations performed by the same examiner in the cited studies was mostly rated as substantial or almost perfect (kappa value ≥ 0.61)^(22,25,33). A lower degree of intra-observer agreement was found in the study by Persichetti *et al.*, with kappa values reported as follows: 0.62 for vascularisation, 0.58 for structure, 0.60 for echogenicity, 0.55 for microcalcifications, 0.54 for macrocalcifications, 0.47 for comet tail artefacts, 0.39 for margins, and 0.35 for shape⁽²⁹⁾.

In summary, inter-observer agreement in the assessment of individual thyroid ultrasound features varies considerably between centres, ranging from slight to almost perfect. The features with the highest disagreement are lesion margins and comet tail artefacts. The level of intra-observer agreement is higher than inter-observer agreement, but still not satisfactory⁽²⁹⁾.

Summary

One method to improve inter-observer agreement involves using a standardised glossary of terms to describe focal lesions on thyroid ultrasound⁽²⁹⁾. Moreover, grading focal thyroid lesions in a structured manner based on dedicated scales/scoring systems (instead of grading individual features) might substantially improve inter-observer agreement^(18,19). The present study highlights the need to diagnose cancers up to 10 mm in diameter, while taking a more liberal approach to biopsy indications in low-risk lesions. On the basis of published studies, sonoelastography has been shown to be a technique that should be

Tab. 3. Comparison of studies published since 2019 assessing inter-observer agreement in the assessment of specific ultrasound imaging features of focal thyroid lesions

	Basha 2019	Dobruch-Sobczak 2019	Itani 2019	Lam 2019	Pang 2019	Persichetti 2020	Phutharak 2019	Seifert 2020	Wildman-Tobriner 2020
Number of nodules assessed	380	20	180	463	189	100	108	80 (40 + 40)	100
Number of researchers	3	5	4	3	2	7	2	4	15
Statistics	Fleiss' κ	Cohen's κ	Cohen's κ	Randolph's κ	Cohen's κ	Cohen's κ	Cohen's κ	Fleiss' κ	Fleiss' κ
Feature on ultrasound examination:									
structure	0.636	0.55	0.43	0.66	0.10–0.64 ³	0.53	0.616	S1: 0.476 S2: 0.674	0.39
echogenicity	0.750	0.48–0.50 ¹	0.252	0.35	0.24–0.53 ⁴	0.47	0.327	S1: 0.440 S2: 0.622	0.39
shape	0.868	–	0.30	–	0.28	0.47	–	S1: 0.537 S2: 0.676	0.38
margins	0.524	0.39	0.23	0.50	0.07–0.14 ⁵	0.33	0.143	S1: 0.431 S2: 0.796	0.18
halo	– ²	0.41	–	–	0.50	–	–	–	–
hyperechogenic foci	0.598	–	–	0.77	–	–	0.288	–	–
microcalcifications	0.957	0.57	0.27	–	0.39	0.47	–	–	0.28
macrocalcifications	0.974	0.61	0.49	–	–	0.38	–	–	0.41
peripheral calcifications	0.604	–	0.39	–	0.33	0.65	–	–	0.26
total calcifications	–	–	0.38	–	–	–	–	S1: 0.405 S2: 0.424	–
comet tail	0.885	–	0.06	–	–	0.11	–	–	0.08
vascularisation	0.211	0.34	–	–	–	0.46	–	–	–
extra-thyroidal infiltration	1.000	0.40	–	0.82	0.24	–	–	–	–

S1 – session 1; S2 – session 2 (conducted after the examiners have discussed all cases from session 1 together)

¹ Feature not assessed in the study.

² Features such as echogenicity compared to thyroid parenchyma ($\kappa = 0.48$), dominant echogenicity compared to thyroid parenchyma ($\kappa = 0.50$), and echogenicity compared to muscle ($\kappa = 0.49$) were evaluated separately.

³ The following features were evaluated separately: solid structure ($\kappa = 0.64$), partially cystic with suspicious features ($\kappa = 0.10$), partially cystic with eccentric solid area ($\kappa = 0.54$), partially cystic without suspicious features ($\kappa = 0.17$), spongiform ($\kappa = 0.62$).

⁴ The following features were assessed separately: nodule significantly hypoechogenic ($\kappa = 0.33$), hypoechogenic ($\kappa = 0.53$), isoechoic ($\kappa = 0.24$), and hyperechogenic ($\kappa = 0.31$).

⁵ The study separately assessed the following features: irregular margins ($\kappa = 0.07$), regular margins ($\kappa = 0.14$).

included in the lexicon of features analysed when deciding to perform a biopsy for focal thyroid lesions. It is also a useful modality for monitoring lesions after FNAB. In the future, genetic testing combined with ultrasound features of focal lesions may contribute to improving diagnostic⁽³⁵⁾ accuracy.

Conflict of interest

The 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.

Author contributions

Original concept of study: KD-S. Writing of manuscript: KD-S, ZA, BM, AS-S, ES-P, KZ, AŽ. Analysis and interpretation of data: KD-S, ZA, BM, AS-S, ES-P, KZ, AŽ. Final acceptance of manuscript: KD-S, ZA, MD, AL, BM, MR, AS-S, ES-P, KZ, AŽ. Collection, recording and/or compilation of data: KD-S, ZA, BM, AS-S, ES-P, KZ, AŽ. Critical review of manuscript: MD, AL, MR.

References

1. Hoang JK, Middleton WD, Tessler FN: Update on ACR TI-RADS: successes, challenges, and future directions, from the *AJR* Special Series on Radiology Reporting and Data Systems. *Am J Roentgenol* 2021; 216: 570–578.
2. Kim SH, Park CS, Jung SL, Kang BJ, Kim JY, Choi JJ *et al.*: Observer variability and the performance between faculties and residents: US Criteria for Benign and Malignant Thyroid Nodules. *Korean J Radiol* 2010; 11: 149–155.
3. Nowicki A, Dobruch-Sobczak K: Introduction to ultrasound elastography. *J Ultrason*. 2016; 16: 113–124.
4. Săftoiu A, Gilja OH, Sidhu PS, Dietrich CF, Cantisani V, Amy D *et al.*: The EFSUMB guidelines and recommendations for the clinical practice of elastography in non-hepatic applications: Update 2018. *Ultraschall Med* 2019; 40: 425–453.
5. Cosgrove D, Barr R, Bojunga J, Cantisani V, Chammas MC, Dighe M *et al.*: WFUMB guidelines and recommendations on the clinical use of ultrasound elastography: part 4. Thyroid. *Ultrasound Med Biol* 2017; 43: 4–26.
6. Jarzab 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.
7. Filho RHC, Pereira FL, Iared W: Diagnostic accuracy evaluation of two-dimensional shear wave elastography in the differentiation between benign and malignant thyroid nodules: systematic review and meta-analysis. *J Ultrasound Med* 2020; 39: 1729–1741.
8. Nattabi HA, Sharif NM, Yahya N, Ahmad R, Mohamad M, Zaki FM *et al.*: Is diagnostic performance of quantitative 2D-shear wave elastography optimal for clinical classification of benign and malignant thyroid nodules? A systematic review and meta-analysis. *Acad Radiol* 2017; S1076-6332(17)30369-0.
9. 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 (Baltimore)* 2017; 96: e8282.
10. Fagin JA, Wells SA: Biologic and clinical perspectives on thyroid cancer. *N Engl J Med* 2016; 375: 1054–1067.
11. Dobruch-Sobczak K, Adamczewski Z, Szczepanek-Parulska E, Migda B, Woliński K, Krauze A *et al.*: Histopathological verification of the diagnostic performance of the EU-TIRADS classification of thyroid nodules—results of a multicenter study performed in a previously iodine-deficient region. *J Clin Med* 2019; 8: E1781.
12. Ha SM, Baek JH, Na DG, Suh CH, Chung SR, Choi YJ *et al.*: Diagnostic performance of practice guidelines for thyroid nodules: thyroid nodule size versus biopsy rates. *Radiology* 2019; 291: 92–99.
13. Zhang T-T, Li C-F, Wen S-S, Huang D-Z, Sun G-H, Zhu Y-X *et al.*: Effects of tumor size on prognosis in differentiated thyroid carcinoma smaller than 2 cm. *Oncol Lett* 2019; 17: 4229–4236.
14. Fleiss JL: Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; 76: 378–382.
15. Hayes AF, Krippendorff K: Answering the call for a standard reliability measure for coding data. *Commun Methods Measures* 2007; 1: 77–89.
16. McGraw KO, Wong SP: Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996; 1: 30–46.
17. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
18. Hoang JK, Middleton WD, Farjat AE, Teeffey SA, Abinanti N, Boschini FJ *et al.*: Interobserver variability of sonographic features used in the American College of Radiology thyroid imaging reporting and data system. *Am J Roentgenol* 2018; 211: 162–167.
19. Grani G, Lamartina L, Cantisani V, Maranghi M, Lucia P, Durante C: Interobserver agreement of various thyroid imaging reporting and data systems. *Endocr Connect* 2018; 7: 1–7.
20. Grani G, Lamartina L, Ascoli V, Bosco D, Nardi F, D'Ambrosio F *et al.*: Ultrasonography scoring systems can rule out malignancy in cytologically indeterminate thyroid nodules. *Endocrine* 2017; 57: 256–261.
21. Koh J, Kim S-Y, Lee HS, Kim E-K, Kwak JY, Moon HJ *et al.*: Diagnostic performances and interobserver agreement according to observer experience: a comparison study using three guidelines for management of thyroid nodules. *Acta Radiol* 2018; 59: 917–923.
22. Lim-Dunham JE, Erdem Toslak I, Alsabban K, Aziz A, Martin B, Okur G *et al.*: Ultrasound risk stratification for malignancy using the 2015 American Thyroid Association Management Guidelines for Children with Thyroid Nodules and Differentiated Thyroid Cancer. *Pediatr Radiol* 2017; 47: 429–436.
23. Sahli ZT, Sharma AK, Canner JK, Karipineni F, Ali O, Kawamoto S *et al.*: TIRADS interobserver variability among indeterminate thyroid nodules: a single institution study. *J Ultrasound Med* 2019; 38: 1807–1813.
24. Basha MAA, Alnaggar AA, Refaat R, El-Maghraby AM, Refaat MM, Abd Elhamed ME *et al.*: The validity and reproducibility of the thyroid imaging reporting and data system (TI-RADS) in categorization of thyroid nodules: multicentre prospective study. *Eur J Radiol* 2019; 117: 184–192.
25. Dobruch-Sobczak K, Migda B, Krauze A, Mlosek K, Slapa RZ, Wareluk P *et al.*: Prospective analysis of inter-observer and intra-observer variability in multi ultrasound descriptor assessment of thyroid nodules. *J Ultrason* 2019; 19: 198–206.
26. Itani M, Assaker R, Moshiri M, Dubinsky TJ, Dighe MK: Inter-observer variability in the American College of Radiology Thyroid Imaging Reporting and Data System: in-depth analysis and areas for improvement. *Ultrasound Med Biol* 2019; 45: 461–470.
27. Lam CA, McGettigan MJ, Thompson ZJ, Khazai L, Chung CH, Centeno BA, *et al.*: Ultrasound characterization for thyroid nodules with indeterminate cytology: inter-observer agreement and impact of combining pattern-based and scoring-based classifications in risk stratification. *Endocrine* 2019; 66: 278–287.
28. Pang Z, Margolis M, Menezes RJ, Maan H, Ghai S: Diagnostic performance of 2015 American Thyroid Association guidelines and inter-observer variability in assigning risk category. *Eur J Radiol Open* 2019; 6: 122–127.
29. Persichetti A, Di Stasio E, Coccaro C, Graziano F, Bianchini A, Di Donna V *et al.*: Inter- and intraobserver agreement in the assessment of thyroid nodule ultrasound features and classification systems: a blinded multicenter study. *Thyroid* 2020; 30: 237–242.
30. Phuttharak W, Boonrod A, Klungboonkron V, Witsawapaisan T: Inter-rater reliability of various thyroid imaging reporting and data system (TIRADS) classifications for differentiating benign from malignant thyroid nodules. *Asian Pac J Cancer Prev* 2019; 20: 1283–1288.
31. Seifert P, Gorges R, Zimny M, Kreissl MC, Schenke S: Interobserver agreement and efficacy of consensus reading in Kwak-, EU-, and ACR-thyroid imaging recording and data systems and ATA guidelines for the ultrasound risk stratification of thyroid nodules. *Endocrine* 2020; 67: 143–154.
32. Wildman-Tobriner B, Ahmed S, Erkanli A, Mazurowski MA, Hoang JK: Using the American College of Radiology Thyroid Imaging Reporting and Data System at the point of care: sonographer performance and interobserver variability. *Ultrasound Med Biol* 2020; 46: 1928–1933.
33. Choi SH, Kim E-K, Kwak JY, Kim MJ, Son EJ: Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* 2010; 20: 167–172.
34. Liu H, Ma A-L, Zhou Y-S, Yang D-H, Ruan J-L, Liu X-D *et al.*: Variability in the interpretation of grey-scale ultrasound features in assessing thyroid nodules: a systematic review and meta-analysis. *Eur J Radiol* 2020; 129: 109050.
35. Lewiński A, Adamczewski Z, Zygmunt A, Markuszewski L, Karbownik-Lewińska M, Stasiak MJ: Correlation between molecular landscape and sonographic image of different variants of papillary thyroid carcinoma. *J Clin Med* 2019; 8: 1916.