RESEARCH

Interobserver agreement of various thyroid imaging reporting and data systems

Giorgio Grani[1], Livia Lamartina[1], Vito Cantisani[2], Marianna Maranghi[1], Piernatale Lucia[1] and **Cosimo Durante[1]**

[1]Department of Internal Medicine and Medical Specialties, 'Sapienza' University of Rome, Rome, Italy
[2]UOS Innovazioni Diagnostiche e Ultrasonografiche, Azienda Ospedaliera Universitaria Policlinico Umberto I, 'Sapienza' University of Rome, Rome, Italy

Correspondence should be addressed to G Grani: giorgio.grani@uniroma1.it

Abstract

Ultrasonography is the best available tool for the initial work-up of thyroid nodules. Substantial interobserver variability has been documented in the recognition and reporting of some of the lesion characteristics. A number of classification systems have been developed to estimate the likelihood of malignancy: several of them have been endorsed by scientific societies, but their reproducibility is yet to be assessed. We evaluated the interobserver variability of the AACE/ACE/AME, ACR, ATA, EU-TIRADS and K-TIRADS classification systems and the interobserver concordance in the indication to FNA biopsy. Two raters independently evaluated 1055 ultrasound images of thyroid nodules identified in 265 patients at multiple time points, in two separate sets (501 and 554 images). After the first set of nodules, a joint reading was performed to reach a consensus in the feature definitions. The interobserver agreement (Krippendorff alpha) in the first set of nodules was 0.47, 0.49, 0.49, 0.61 and 0.53, for AACE/ACE/AME, ACR, ATA, EU-TIRADS and K-TIRADS systems, respectively. The agreement for the indication to biopsy was substantial to near-perfect, being 0.73, 0.61, 0.75, 0.68 and 0.82, respectively (Cohen's kappa). For all systems, agreement on the nodules of the second set increased. Despite the wide variability in the description of single ultrasonographic features, the classification systems may improve the interobserver agreement that further ameliorates after a specific training. When selecting nodules to be submitted to FNA biopsy, that is main purpose of these classifications, the interobserver agreement is substantial to almost perfect.

Key Words
- TIRADS
- thyroid nodule
- interobserver variability
- agreement
- reliability

Endocrine Connections (2018) **7**, 1–7

Introduction

Thyroid nodules are an increasingly common finding during imaging examinations of the neck, but only a small proportion of these lesions ultimately prove to be malignant (1). Fine-needle aspiration (FNA) biopsy plays a major role the differential diagnosis, but its execution needs to be selective (2), due to the associated costs, potential non-diagnostic results (3), and the risk of overdiagnosis (4). Ultrasonography (US) is currently the best diagnostic tool available for the initial work-up of thyroid nodules. Certain US features are more or less strongly associated with nodule malignancy (5). However, the diagnostic accuracy of each single feature is limited, and substantial interobserver variation that has been documented in the recognition and reporting of some of the lesion characteristics. The low reproducibility of US classifications of thyroid nodules is aggravated by the heterogenous professional profiles and experience levels of the individuals who perform thyroid US scans (e.g., technicians, radiologists, clinicians) and interpret the results (6).

A number of classification systems have been developed that use composite patterns of US findings to

http://www.endocrineconnections.org
https://doi.org/10.1530/EC-17-0336

estimate the likelihood of malignancy and, in particular, to identify nodules that need to be scheduled for FNA biopsy (7, 8). Several of these systems have been endorsed by international scientific bodies (9, 10, 11, 12, 13), but their reproducibility is yet to be assessed.

We conducted a retrospective analysis of recorded US images to evaluate the interobserver variability in the application the main thyroid nodule US classification systems and the resulting interobserver concordance in the indication to perform FNA biopsy.

Subjects and methods

Cases

We conducted a retrospective analysis of 1055 ultrasound images of thyroid nodules identified in 265 patients (each with less than four nodules). All had originally been classified as benign (those with suspicious US features but benign cytology) or presumably benign (nodules with no suspicious ultrasound features) and managed with active surveillance as long as there was no evidence of malignancy. The images had been acquired in our thyroid cancer unit at the time of nodule detection and/or during the first five years thereafter and stored in order to precisely document the main nodule features over time. All examinations were performed by a single examiner, with 10-year experience in thyroid US, using an Esaote MyLab 25 (Esaote SpA, Genoa, Italy) ultrasound system with a high-frequency linear transducer. The cases included in the present study represent our center's contribution to a larger multicenter cohort of patients analyzed prospectively to explore the natural history of benign thyroid nodules. The original study, which has been described in detail elsewhere, (14, 15) was conducted with Institutional Review Board (Sapienza University of Rome Ethics Committee) approval and the written informed consent of all participants. Between January 1, 2006, and January 31, 2008, centers enrolled a consecutive series of euthyroid patients presenting with one to four asymptomatic thyroid nodules, measuring 4–40 mm, that were presumed to be benign on the basis of a benign cytology report or the absence of suspicious sonographic features. Exclusion criteria were levothyroxine therapy during the study or in the 6 months preceding enrollment, history of surgical or nonsurgical thyroid interventions, history of thyroid autoimmunity and subacute thyroiditis. Enrolled patients were followed with yearly clinical and US evaluations.

For the purposes of the present study, the images of the 1055 nodules (at various follow-up points) were converted to and stored as deidentified bitmap (BMP) files. The stored files were then randomly divided into two groups: set 1 (501 nodules) and set 2 (554 nodules).

Review of ultrasound images

Analysis of image set 1

The BMP file of each nodule in set 1 was independently reviewed on a single liquid crystal display monitor by two clinicians (GG and LL). Each reader had 6 years of experience in thyroid US imaging, although they had been trained in two different thyroid units. The readers were blinded to the identity of the patient, the date of the scan and all other clinical information regarding the case.

Using standardized in-house multiple-choice forms developed on the basis of published suggestions (6, 16), the two readers rated the following US features of each nodule: margins (well-defined, ill-defined, microlobulated or irregular, infiltrative, peripheral halo); composition (solid, cystic, mixed); echogenicity (hyperechoic, isoechoic, hypoechoic—all with respect to the surrounding thyroid parenchyma—or markedly hypoechoic, i.e., less echoic than the adjacent strap muscle); calcifications (absent, microscopic, macroscopic—the latter including eggshell calcifications); other hyperechoic foci (comet-tail artifacts or indeterminate, the latter including areas of fibrosis). For mixed-content nodules, the reader also rated the location of the solid component (diffuse-not nodular, peripheral, central), and recorded the spongiform appearance. Echogenicity and structure were not evaluated in nodules with complete rim calcification. Three other parameters were not rated by either of the readers: (1) nodule diameters (transverse, anteroposterior and longitudinal), which had been recorded during the original scan and were visible on the stored images; (2) nodule shape, parallel or non-parallel (or taller-than-wide), the classification of which is strictly dependent on the nodule dimensions and (3) vascularity, because it was not available in all of the cases. These features were excluded from analyses of interobserver agreement.

For each nodule, the ratings of each reader (together with those recorded during the original examination for nodule size and shape) were elaborated automatically using an in-house algorithm to classify the nodule according to the following five systems: Guidelines of the American Association of Clinical Endocrinologists/American College of Endocrinology/Associazione Medici Endocrinologi (AACE/ACE/AME) (9); the TIRADS system

http://www.endocrineconnections.org
https://doi.org/10.1530/EC-17-0336

developed by the American College of Radiologists (ACR) (10); the 2015 Guidelines of the American Thyroid Association (ATA) (11); the EU-TIRADS system proposed by the European Thyroid Association (12) and the Korean Society of Thyroid Radiology's K-TIRADS system (13).

Training session

Two weeks after their independent reviews and classification of set 1 images, the two readers jointly reviewed the results and the images of all 501 nodules. Discrepancies between their ratings were discussed and a consensus decision reached for each nodule feature. The consensus ratings were then elaborated using the same algorithm to generate new risk classifications of the nodules for each system.

Analysis of image set 2

Four weeks after completion of the training session, the two readers were asked to independently review US images of the 554 nodules of set 2 and to rate them using the same methods employed for set 1. The results were automatically elaborated to obtain risk classifications for each nodule according to the same five systems.

Analysis of data

For each set of nodules, we assessed inter-reader agreement at the level of single features of the nodule, risk-class assignment based on each of the five US classification systems and the advisability/non-advisability of FNA biopsy based on the risk-class assignments. Agreement on ordinal ratings was assessed with the Krippendorff α statistic (17). Values close to 1 indicate high inter-reader agreement, and values above 0.65 are considered an acceptable basis for tentative conclusions. Interobserver agreement on nominal, dichotomic ratings was evaluated using Cohen's kappa statistic. Values less than 0.20 are considered indicative of slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement and 0.81–1.00, near-perfect agreement (18). For all statistics, 95% confidence intervals (CI) were also calculated. All analyses were performed with IBM SPSS Statistics program, v. 23.0 (IBM).

Results

Supplementary Table 1 (see section on supplementary data given at the end of this article) shows the estimated risks of malignancy for set 1 and set 2 nodules, respectively, based on the consensus judgments of the two readers. (These data are presented for informative purposes only. All of the analyses presented below were restricted to assessment of agreement between the independent ratings of the readers in their descriptions of the nodules.)

Table 1 shows the interobserver agreement for the recognition of single US features in set 1 and set 2, as well as those reported in previous studies of this type (19, 20, 21, 22, 23, 24, 25, 26, 27, 28). Agreement between the readers in our study was highest regarding the presence/absence of calcifications (particularly macrocalcifications). Alpha values for single features in the second set of nodules were not significantly different from those for set 1.

Table 2 summarizes the data on interobserver agreement nodule risk classification for the five reporting systems tested. For each nodule set, results are presented for all lesions and for the subset of lesions exceeding 1 cm in diameter. For all five systems, agreement on the nodules of set 2 was appreciably better than that recorded for the set 1. This effect was independent of nodule size, but the degree of improvement varied from system to system. In all of the comparisons, classification according to the EU-TIRADS system displayed the highest level of reproducibility.

As shown in Table 3, agreement in the identification of set 1 nodules that required FNA biopsy according to the five systems was already substantial to near-perfect. Nevertheless, the Cohen kappas for the second set of nodules were consistently higher for all five classifications systems.

Discussion

Several classification systems have been developed to help physicians decide which of the myriad thyroid nodules discovered each year need to be evaluated with FNA biopsy. These efforts were motivated by the conviction that defining the risk of malignancy based on the presence/absence of composite sets of US features would reduce the interobserver variability seen when risk estimates were based on the presence/absence of individual features. However, all are heavily dependent on the operator's ability to accurately describe key nodule features, such as composition or echogenicity (30), which, in previous studies, have been characterized by suboptimal interobserver agreement (28, 31). The aim of the present study was evaluating the interobserver reliability in the application of the main thyroid nodule US classification systems. Few attempts have been made to demonstrate whether or not these systems do allow a more reproducible

http://www.endocrineconnections.org
https://doi.org/10.1530/EC-17-0336

Table 1 Interobserver agreement for single US features, compared with published data.

	This study		Choi, 2010 (22)[a]	Kim, 2012 (23)	Kim, 2010 (24)[a]		
Nodules	Set 1: 501	Set 2: 554	204	80	133		
Statistics	Krippendorff alpha		Kappa	Kappa	Kappa		
Observers	2 clinicians, same level of experience		4 exp. radiologists, same institution	7 resident radiologists, 2 different units	9	5 faculty radiologists	4 residents
Echogenicity	0.56 (0.46-–0.66)	0.66 (0.59-–0.73)	0.45	0.5	0.46	0.57	0.34
Composition	0.52 (0.34–0.68)	0.5 (0.29–0.68)	0.59	0.48	0.36	0.64	0.18
Shape	N/A	N/A	0.61	0.57	0.4	0.46	0.34
Margin	0.51 (0.43–0.58)	0.44 (0.34–0.53)	0.61	0.49	0.25	0.4	0.19
Vascularity	N/A	N/A	0.46	N/A	N/A	N/A	N/A
Calcification	0.8 (0.63–0.93)	0.89 (0.75–1)	0.58	0.62	0.47	0.63	0.42
Micro-	0.49 (−0.28–1)	0.39 (−0.49–1)	0.51	0.59	N/A	N/A	N/A
Macro-	0.85 (0.59–1)	0.83 (0.6–1)	0.39	0.39	N/A	N/A	N/A
Echogenic foci	0.48 (0.3–0.64)	0.35 (0.17–0.52)	N/A	N/A	N/A	N/A	N/A
Capsule invasion	0.11 (−0.91–1)	0.4 (−1–1)	N/A	N/A	N/A	N/A	N/A

[a]Maximum *k* value reported in any of the two sessions; [b]All features were grouped in two classes.
AUS/FLUS, Atypia of Undetermined Significance or Follicular Lesion of Undetermined Significance; FN, Follicular Neoplasm; N/A, not available.

stratification of thyroid nodules. On the basis of the data collected in the present study, the answer appears to be 'yes.' On the whole, classification of thyroid nodules using any of the five systems we tested is associated with higher interobserver agreement than classification based on single suspicious features, although there is clearly room for further improvement. More importantly, identification of nodules that require FNA biopsy based on these classification systems, which is in fact their main purpose, is associated with substantial to near-perfect agreement.

Other factors, aside from the experience in general thyroid imaging and nodule size, should be taken into account when working to improve reliability of standard US reporting of thyroid nodules. Several approaches have already been proposed to achieve this target, such as quantitative evaluation of echogenicity (31) and more systematic reporting schemas. (6, 12, 32).

A specific 'hands-on' training involving joint reading of US images can improve the reproducibility of all the risk classifications, even for trained readers with similar levels of experience, with significant improvements for ATA, K-TIRADS and EU-TIRADS systems, even if no significant difference is recorded for any single variable. Some of the classification systems place substantial weight on certain individual features (in most cases, solid structure and hypoechogenicity), and the reproducibility of these classifications depends on the agreement between readers in describing these particular features. However, the weight of each feature varies across various systems (echogenicity being, for example, the most important

Table 2 Interobserver agreement for nodule classification the various US classification systems endorsed by scientific societies.

	All nodules		Only nodules >1 cm	
	Set 1 (*n* = 501)	Set 2 (*n* = 554)	Set 1 (*n* = 219)	Set 2 (*n* = 207)
AACE/ACE/AME	0.47 (0.35–0.57)	0.61 (0.49–0.72)	0.53 (0.41–0.63)	0.58 (0.44–0.71)
ACR TIRADS	0.49 (0.4–0.57)	0.57 (0.5–0.63)	0.45 (0.31–0.58)	0.62 (0.51–0.71)
ATA	0.49 (0.41–0.57)	0.65 (0.58–0.71)	0.44 (0.3–0.56)	0.72 (0.62–0.81)
EU-TIRADS	0.61 (0.54–0.68)	0.75 (0.69–0.81)	0.63 (0.52–0.72)	0.77 (0.71–0.83)
K-TIRADS	0.53 (0.43–0.62)	0.66 (0.57–0.73)	0.54 (0.43–0.66)	0.74 (0.6–0.86)

Krippendorff alpha (95% CI).
AACE/ACE/AME, American Association of Clinical Endocrinologists/American College of Endocrinology/Associazione Medici Endocrinologi; ACR, American College of Radiologists; ATA, American Thyroid Association; EU-TIRADS, European Thyroid Imaging Reporting and Data Systems; K-TIRADS, Korean Thyroid Imaging Reporting and Data Systems.

http://www.endocrineconnections.org
https://doi.org/10.1530/EC-17-0336

Koltin, 2016 (25)	Lim-Dunham, 2017 (26)	Norlen, 2014 (32)[b]	Park, 2009 (27)	Park, 2010 (28)	Park, 2012 (29)	Wienke, 2003 (30)	Grani, 2017 (31)
27 (pediatric pop.)	39 (pediatric population)	141 (AUS/FLUS)	52 (all malignant)	133	400	70	49 (AUS/FLUS/FN)
Kappa	Kappa	Kappa	Spearman corr.	Kappa	Kappa	Kappa	Krippendorff alpha
3 exp. radiologists	2 exp. radiologists	2 trained surgeons	3 radiologists (1–7 year exp.)	5 radiologists (1-–6 6 year exp.)	3 radiologists (7-–10 10 year exp.), different institution	2 radiologists	3 clinicians, 2 different units
0.46	0.54 (0.37-–0.73)	0.94	0.04–0.45	0.57	0.504	0.37	0.58 (0.45–0.70)
N/A	0.8 (0.53–0.99)	N/A	0.70–1.00	0.64	0.818	0.62	0.5 (0.29–0.68)
N/A	0.29 (0.01–0.72)	0.61	0.48–0.79	0.42	0.42	N/A	N/A
0.58	0.6 (0.4–0.79)	0.6	0.03–0.29	0.34	0.33	0.13	0.57 (0.43–0.70)
0.18	0.76 (0.52–0.99)	0.74	N/A	N/A	N/A	0.75	N/A
N/A	N/A	N/A	0.47–0.62	0.55	0.479	0.91	0.68 (0.46–0.88)
0.59	N/A	0.79	N/A	0.54	N/A	N/A	N/A
N/A	N/A	N/A	N/A	0.4	N/A	N/A	N/A
N/A	0.77 (0.56–0.99)	N/A	N/A	N/A	N/A	N/A	0.34 (0.12–0.56)
N/A	N/A	N/A	N/A	0.32	N/A	N/A	0.32 (–0.61–1.0)

one for EU-TIRADS, and composition for K-TIRADS). Furthermore, the precise identification of the full range of possible descriptions is not needed for classification: for example, only irregular or infiltrative margins impact the nodule classification, while the difference between regular, ill-defined and hypoechoic halo is negligible. The differences in the system structure may also explain the higher interobserver agreement for some systems. Classifications with a lower number of high-suspicious features and more gradual scoring show better interobserver reliability (such as EU-TIRADS) than the ones in which a single description may dramatically change the nodule classification (ACR TIRADS, ATA guidelines). Similarly, the agreement on the FNA biopsy indication is influenced by the philosophy behind each classification: K-TIRADS system has the highest reproducibility because the vast majority of low- and intermediate-suspicion nodules are submitted to FNA biopsy according to the Korean system, while the other classification are more conservative, with higher dimensional cutoff and deeper management differences between low- and intermediate-suspicion categories – that represent a source of variability.

Table 3 Interobserver agreement[a] on indications for FNA biopsy according to the various guidelines.

	Set 1 ($n=501$)	**Set 2** ($n=554$)
AACE/ACE/AME	0.73 (0.64–0.82)	0.82 (0.75–0.89)
ACR TIRADS	0.61 (0.5–0.72)	0.73 (0.63–0.82)
ATA	0.75 (0.67–0.82)	0.82 (0.75–0.89)
EU-TIRADS	0.68 (0.58–0.79)	0.74 (0.65–0.83)
K-TIRADS	0.82 (0.76–0.88)	0.91 (0.86–0.95)

[a]Cohen kappa (95% confidence intervals).
AACE/ACE/AME, American Association of Clinical Endocrinologists/American College of Endocrinology/Associazione Medici Endocrinologi; ACR, American College of Radiologists; ATA, American Thyroid Association; EU-TIRADS, European Thyroid Imaging Reporting and Data Systems; K-TIRADS, Korean Thyroid Imaging Reporting and Data Systems.

These findings are similar to that reported by Cheng for the first TIRADS ever proposed ($k=0.61$, moderate to substantial agreement) (33), although in other hands, the agreement observed with this system was appreciably lower ($k=0.27$) (34). Our group reported, in a previous study (28), a suboptimal agreement in a small series of cytologically indeterminate nodules (Krippendorff alpha 0.36 for ATA classification and 0.42 for TIRADS classification proposed by Kwak (35)). An even higher k value of 0.72 was reported by Russ and coworkers for the system they proposed in 2013 (36), which is the basis of the current EU-TIRADS systems. According to some evidence, thyroid computer-aided diagnosis (CAD) using artificial intelligence may further improve diagnosis reliability. It was reported that the use of thyroid CAD to differentiate malignant from benign nodules showed accuracy similar to that obtained by radiologists (37, 38) and may reduce intra- and interobserver variability.

Our study has some obvious limitations. First, owing to its retrospective nature, it was not possible to consider

http://www.endocrineconnections.org
https://doi.org/10.1530/EC-17-0336

the influence on interobserver agreement of scan-related variables, such as probe inclination, US equipment used, operating conditions and setting. Second, the sample consisted exclusively of nodules that had initially been classified as benign. This could potentially overestimate the agreement. However, all categories are represented in this series, and the sample size is broader than that in other non-selected series. Third, the readers involved in this study were 'peers' in terms of experience as well as staff members of in the same thyroid cancer unit, both of which factors could conceivably have contributed to the improved interobserver agreement we observed after the training session. Similar initiatives in larger, more heterogeneous groups might not have the same results. Furthermore, this study does not provide data about increasingly used ancillary techniques, like elastosonography (39, 40).

In conclusion, despite the wide variability in the description of single US features, the US classification systems may improve the interobserver agreement, that further ameliorates after a specific training. When selecting nodules to be submitted to FNA biopsy, that is main purpose of these classifications, the interobserver agreement is substantial to almost perfect.

Supplementary data

This is linked to the online version of the paper at https://doi.org/10.1530/EC-17-0336.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

Acknowledgements

G G and L L contributed to this paper as recipient of the PhD program of Biotechnologies and Clinical Medicine of the University of Rome, Sapienza. Writing support was provided by Marian Everett Kent, BSN.

References

1 Filetti S, Durante C & Torlontano M. Nonsurgical approaches to the management of thyroid nodules. *Nature Clinical Practice: Endocrinology and Metabolism* 2006 **2** 384–394. (https://doi.org/10.1038/ncpendmet0215)

2 Lamartina L, Deandreis D, Durante C & Filetti S. ENDOCRINE TUMOURS: imaging in the follow-up of differentiated thyroid cancer: current evidence and future perspectives for a risk-adapted approach. *European Journal of Endocrinology* 2016 **175** R185–R202. (https://doi.org/10.1530/EJE-16-0088)

3 Grani G, Calvanese A, Carbotta G, D'Alessandri M, Nesca A, Bianchini M, Del Sordo M & Fumarola A. Intrinsic factors affecting adequacy of thyroid nodule fine-needle aspiration cytology. *Clinical Endocrinology* 2013 **78** 141–144. (https://doi.org/10.1111/j.1365-2265.2012.04507.x)

4 Brito JP, Morris JC & Montori VM. Thyroid cancer: zealous imaging has increased detection and treatment of low risk tumours. *BMJ* 2013 **347** f4706. (https://doi.org/10.1136/bmj.f4706)

5 Brito JP, Gionfriddo MR, Al Nofal A, Boehmer KR, Leppin AL, Reading C, Callstrom M, Elraiyah TA, Prokop LJ, Stan MN, *et al*. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. *Journal of Clinical Endocrinology and Metabolism* 2014 **99** 1253–1263. (https://doi.org/10.1210/jc.2013-2928)

6 Su HK, Dos Reis LL, Lupo MA, Milas M, Orloff LA, Langer JE, Brett EM, Kazam E, Lee SL, Minkowitz G, *et al*. Striving toward standardization of reporting of ultrasound features of thyroid nodules and lymph nodes: a multidisciplinary consensus statement. *Thyroid* 2014 **24** 1341–1349. (https://doi.org/10.1089/thy.2014.0110)

7 Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A & Dominguez M. An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. *Journal of Clinical Endocrinology and Metabolism* 2009 **94** 1748–1751. (https://doi.org/10.1210/jc.2008-1724)

8 Ozel A, Erturk SM, Ercan A, Yılmaz B, Basak T, Cantisani V, Basak M & Karpat Z. The diagnostic efficiency of ultrasound in characterization for thyroid nodules: how many criteria are required to predict malignancy? *Medical Ultrasonography* 2012 **14** 24–28.

9 Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L, Paschke R, Valcavi R & Vitti P. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules – 2016 update. *Endocrine Practice* 2016 **22** 622–639. (https://doi.org/10.4158/EP161208.GL)

10 Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, *et al*. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *Journal of the American College of Radiology* 2017 **14** 587–595. (https://doi.org/10.1016/j.jacr.2017.01.046)

11 Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, *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. (https://doi.org/10.1089/thy.2015.0020)

12 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. *European Thyroid Journal* 2017 **6** 225–237. (https://doi.org/10.1159/000478927)

13 Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, Lim HK, Moon WJ, Na DG, Park JS, *et al*. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of thyroid radiology consensus statement and recommendations. *Korean Journal of Radiology* 2016 **17** 370–395. (https://doi.org/10.3348/kjr.2016.17.3.370)

14 Durante C, Costante G, Lucisano G, Bruno R, Meringolo D, Paciaroni A, Puxeddu E, Torlontano M, Tumino S, Attard M, *et al*. The natural history of benign thyroid nodules. *JAMA* 2015 **313** 926–935. (https://doi.org/10.1001/jama.2015.0956)

15 Grani G, Bruno R, Lucisano G, Costante G, Meringolo D, Puxeddu E, Torlontano M, Tumino S, Attard M, Lamartina L, *et al*. Temporal changes in thyroid nodule volume: lack of effect

on paranodular thyroid tissue volume. *Thyroid* 2017 **27** 1378–1384. (https://doi.org/10.1089/thy.2017.0201)
16 Grant EG, Tessler FN, Hoang JK, Langer JE, Beland MD, Berland LL, Cronan JJ, Desser TS, Frates MC, Hamper UM, *et al*. Thyroid ultrasound reporting Lexicon: white paper of the ACR thyroid imaging, reporting and data system (TIRADS) committee. *Journal of the American College of Radiology* 2015 **12** 1272–1279. (https://doi.org/10.1016/j.jacr.2015.07.011)
17 Hayes AF & Krippendorff K. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 2007 **1** 77–89. (https://doi.org/10.1080/19312450709336664)
18 Landis JR & Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 **33** 159–174. (https://doi.org/10.2307/2529310)
19 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. (https://doi.org/10.1089/thy.2008.0354)
20 Kim HG, Kwak JY, Kim EK, Choi SH & Moon HJ. Man to man training: can it help improve the diagnostic performances and interobserver variabilities of thyroid ultrasonography in residents? *European Journal of Radiology* 2012 **81** e352–e356. (https://doi.org/10.1016/j.ejrad.2011.11.011)
21 Kim SH, Park CS, Jung SL, Kang BJ, Kim JY, Choi JJ, Kim YI, Oh JK, Oh JS, Kim H, *et al*. Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules. *Korean Journal of Radiology* 2010 **11** 149–155. (https://doi.org/10.3348/kjr.2010.11.2.149)
22 Koltin D, O'Gorman CS, Murphy A, Ngan B, Daneman A, Navarro OM, Garcia C, Atenafu EG, Wasserman JD, Hamilton J, *et al*. Pediatric thyroid nodules: ultrasonographic characteristics and inter-observer variability in prediction of malignancy. *Journal of Pediatric Endocrinology and Metabolism* 2016 **29** 789–794. (https://doi.org/10.1515/jpem-2015-0242)
23 Lim-Dunham JE, Erdem Toslak I, Alsabban K, Aziz A, Martin B, Okur G & Longo KC. Ultrasound risk stratification for malignancy using the 2015 American Thyroid Association Management Guidelines for children with thyroid nodules and differentiated thyroid cancer. *Pediatric Radiology* 2017 **47** 429–436. (https://doi.org/10.1007/s00247-017-3780-6)
24 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. *American Journal of Roentgenology* 2009 **193** W416–W423. (https://doi.org/10.2214/AJR.09.2541)
25 Park CS, Kim SH, Jung SL, Kang BJ, Kim JY, Choi JJ, Sung MS, Yim HW & Jeong SH. Observer variability in the sonographic evaluation of thyroid nodules. *Journal of Clinical Ultrasound* 2010 **38** 287–293. (https://doi.org/10.1002/jcu.20689)
26 Park SJ, Park SH, Choi YJ, Kim DW, Son EJ, Lee HS, Yoon JH, Kim EK, Moon HJ & Kwak JY. Interobserver variability and diagnostic performance in US assessment of thyroid nodule according to size. *Ultraschall in Der Medizin* 2012 **33** E186–E190. (https://doi.org/10.1055/s-0032-1325404)
27 Wienke JR, Chong WK, Fielding JR, Zou KH & Mittelstaedt CA. Sonographic features of benign thyroid nodules: interobserver reliability and overlap with malignancy. *Journal of Ultrasound in Medicine* 2003 **22** 1027–1031. (https://doi.org/10.7863/jum.2003.22.10.1027)
28 Grani G, Lamartina L, Ascoli V, Bosco D, Nardi F, D'Ambrosio F, Rubini A, Giacomelli L, Biffoni M, Filetti S, *et al*. Ultrasonography scoring systems can rule out malignancy in cytologically indeterminate thyroid nodules. *Endocrine* 2017 **57** 256–261. (https://doi.org/10.1007/s12020-016-1148-6)
29 Norlen O, Popadich A, Kruijff S, Gill AJ, Sarkis LM, Delbridge L, Sywak M & Sidhu S. Bethesda III thyroid nodules: the role of ultrasound in clinical decision making. *Annals of Surgical Oncology* 2014 **21** 3528–3533. (https://doi.org/10.1245/s10434-014-3749-8)
30 Na DG, Baek JH, Sung JY, Kim JH, Kim JK, Choi YJ & Seo H. Thyroid imaging reporting and data system risk stratification of thyroid nodules: categorization based on solidity and echogenicity. *Thyroid* 2016 **26** 562–572. (https://doi.org/10.1089/thy.2015.0460)
31 Grani G, D'Alessandri M, Carbotta G, Nesca A, Del Sordo M, Alessandrini S, Coccaro C, Rendina R, Bianchini M, Prinzi N, *et al*. Grey-scale analysis improves the ultrasonographic evaluation of thyroid nodules. *Medicine* 2015 **94** e1129. (https://doi.org/10.1097/MD.0000000000001129)
32 Leenhardt L, Erdogan MF, Hegedus L, Mandel SJ, Paschke R, Rago T & Russ G. 2013 European Thyroid Association Guidelines for cervical ultrasound scan and ultrasound-guided techniques in the postoperative management of patients with thyroid cancer. *European Thyroid Journal* 2013 **2** 147–159. (https://doi.org/10.1159/000354537)
33 Cheng SP, Lee JJ, Lin JL, Chuang SM, Chien MN & Liu CL. Characterization of thyroid nodules using the proposed thyroid imaging reporting and data system (TI-RADS). *Head and Neck* 2013 **35** 541–547. (https://doi.org/10.1002/hed.22985)
34 Friedrich-Rust M, Meyer G, Dauth N, Berner C, Bogdanou D, Herrmann E, Zeuzem S & Bojunga J. Interobserver agreement of Thyroid Imaging Reporting and Data System (TIRADS) and strain elastography for the assessment of thyroid nodules. *PLoS ONE* 2013 **8** e77927. (https://doi.org/10.1371/journal.pone.0077927)
35 Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, Jung HK, Choi JS, Kim BM & Kim EK. 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. (https://doi.org/10.1148/radiol.11110206)
36 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. *European Journal of Endocrinology* 2013 **168** 649–655. (https://doi.org/10.1530/EJE-12-0936)
37 Chang Y, Paul AK, Kim N, Baek JH, Choi YJ, Ha EJ, Lee KD, Lee HS & Shin D. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. *Medical Physics* 2016 **43** 554. (https://doi.org/10.1118/1.4939060)
38 Choi YJ, Baek JH, Park HS, Shim WH, Kim TY, Shong YK & Lee JH. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. *Thyroid* 2017 **27** 546–552. (https://doi.org/10.1089/thy.2016.0372)
39 Cantisani V, Grazhdani H, Drakonaki E, D'Andrea V, Di Segni M, Kaleshi E, Calliada F, Catalano C, Redler A, Brunese L, *et al*. Strain US elastography for the characterization of thyroid nodules: advantages and limitation. *International Journal of Endocrinology* 2015 **2015** 908575. (https://doi.org/10.1155/2015/908575)
40 Cosgrove D, Barr R, Bojunga J, Cantisani V, Chammas MC, Dighe M, Vinayak S, Xu JM & Dietrich CF. WFUMB guidelines and recommendations on the clinical use of ultrasound elastography: part 4. Thyroid. *Ultrasound in Medicine and Biology* 2017 **43** 4–26. (https://doi.org/10.1016/j.ultrasmedbio.2016.06.022)

Received in final form 4 November 2017
Accepted 9 November 2017
Accepted preprint published online 13 November 2017