Endocrine (2019) 66:278–287
https://doi.org/10.1007/s12020-019-02000-0

ORIGINAL ARTICLE

Check for updates

Ultrasound characterization for thyroid nodules with indeterminate cytology: inter-observer agreement and impact of combining pattern-based and scoring-based classifications in risk stratification

Cesar A. Lam[1] · Melissa J. McGettigan[1] · Zachary J. Thompson[2] · Laila Khazai[3] · Christine H. Chung[4] · Barbara A. Centeno[3] · Bryan McIver[4] · Pablo Valderrabano[4,5]

Received: 28 March 2019 / Accepted: 29 June 2019 / Published online: 12 July 2019

Abstract

Background The American Thyroid Association (ATA) sonographic patterns stratify the risk of malignancy of cytologically indeterminate thyroid nodules (ITNs). This study aimed to (1) assess inter-observer agreement for sonographic features and patterns; (2) identify potential sources of disagreement; and (3) evaluate whether the number of suspicious features risk-stratifies non-ATA and high-suspicion patterns.

Methods Three observers independently reviewed the ultrasound images of 463 ITNs with histological follow-up consecutively evaluated between October 2008 and June 2015 at an academic cancer center. Each observer evaluated individual sonographic features. ATA sonographic patterns were derived from the interpretation of sonographic features. Nodules not fitting into any of the proposed patterns were clustered into a non-ATA pattern.

Results The inter-observer agreement for ATA sonographic patterns and echogenicity was fair, moderate for margins, good for composition and echogenic foci, and very good for extrathyroidal extension and lymph node metastasis. The interpretation of each sonographic feature was significantly different between observers, and there was complete disagreement in at least one of the features in 104 (22%) nodules. A total of 169 nodules (37%) were classified into the non-ATA pattern. The number of suspicious features allowed risk stratifying nodules with non-ATA and high-suspicion sonographic patterns. Most Non-invasive Follicular Thyroid Neoplasms with Papillary-like Nuclear Features had 0–1 suspicious features and none had >2.

Conclusions Echogenicity interpretation was the greatest source of disagreement. The number of suspicious features risk-stratifies ITNs with non-ATA or high-suspicion patterns. Future studies attempting to objectivize the interpretation of echogenicity and heterogeneity are needed.

Keywords Thyroid ultrasound · Thyroid cytology · Thyroid nodules · Thyroid cancer · Non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP)

Supplementary information The online version of this article (https://doi.org/10.1007/s12020-019-02000-0) contains supplementary material, which is available to authorized users.

✉ Cesar A. Lam
cesar.lam@moffitt.org

[1] Department of Diagnostic Imaging, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, USA

[2] Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, USA

[3] Department of Anatomic Pathology, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, USA

[4] Department of Head and Neck-Endocrine Oncology, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, USA

[5] Present address: Department of Endocrinology and Nutrition, Hospital Universitario Ramón y Cajal, IRYCIS, Ctra. de Colmenar Viejo km. 9,100, 28034 Madrid, Spain

Introduction

The widespread use of diagnostic imaging tests in the last decades has been recognized as the most significant factor leading to the current thyroid cancer epidemic [1–5]. Diagnostic imaging, however, is not just uncovering thyroid cancers but also a much larger reservoir of asymptomatic benign thyroid nodules. Because most thyroid cancers present as thyroid nodules, the identification of a thyroid nodule often triggers a cascade of diagnostic tests, which are also on the rise [5, 6].

The American Thyroid Association guidelines for the management of patients with thyroid nodules and differentiated thyroid cancer (ATA guidelines) recommend performing a thyroid ultrasound in every patient with a known or suspected thyroid nodule to characterize its sonographic features [7]. Several sonographic features have been consistently associated with an increased risk of malignancy, including solid composition, hypoechogenicity, irregular or infiltrative margins, microcalcifications, and a taller than wide shape [8–10]. Unfortunately, due to subjective interpretation of ultrasound images, the inter-observer agreement for these features is moderate [11–13]. Furthermore, none of these features are sensitive or specific enough to be used in isolation [8–10]. For this reason, several classification systems have tried to integrate the sonographic features into different groups or patterns that stratify the risk of cancer and guide the need for biopsy. The 2015 ATA guidelines proposed a new classification that is based on the recognition of five different sonographic patterns: benign, very-low, low, intermediate, and high suspicion. Each of these patterns is associated with an estimated risk of malignancy (<1%, <3%, 5–10%, 10–20%, and >70–90%, respectively) and a recommended size threshold for biopsy (no biopsy, ≥2 cm, ≥1.5 cm, ≥1 cm, and ≥1 cm, respectively), which is based on two factors, the estimated probability of cancer and the probability of that cancer to be clinically relevant [7].

If a thyroid nodule meets the size threshold for the specific sonographic pattern, a fine needle aspiration is usually performed and the cytological interpretation is used to decide the clinical management [7]. If it is benign, the patient's nodule is usually followed and if it is malignant, it is often resected. Unfortunately, 25% of the biopsies render an indeterminate cytologic diagnosis for a wide variety of reasons, although they are usually clustered for management recommendations. We have recently shown that the rate of malignancy of these nodules can be stratified according to the sonographic pattern [14]. Whereas the risk stratification was very consistent, there was a poor correlation for the distribution of the sonographic patterns between three observers. Furthermore, over one third of the nodules were found to have a sonographic pattern not described in the ATA classification, which were grouped -into a "non-ATA" pattern. Unlike nodules in the very-low, low, or intermediate suspicion sonographic patterns; nodules in the non-ATA or high-suspicion sonographic pattern cluster a constellation of nodules that exhibit a variable number of suspicious sonographic features. We hypothesized that the number of suspicious features could help risk-stratify these two categories further. In this study, we aimed to (1) assess the inter-observer agreement for individual sonographic features in the context of a large cohort of thyroid nodules with indeterminate cytology; (2) identify potential sources of disagreement; and (3) evaluate the impact of applying a scoring system to risk-stratify non-ATA and high-suspicion ATA sonographic patterns.

Methods

Three observers independently reviewed the presurgical thyroid ultrasound images of all cytologically indeterminate thyroid nodules (ITNs; atypia/follicular lesion of undetermined significance [AUS/FLUS] and follicular neoplasm [FN]) consecutively evaluated at an academic cancer center between October 2008 and April 2015, which had histological follow-up before June 2015. Each observer (two radiologists [with 12 and 5 years of ultrasound experience] and one endocrinologist [with 8 years of ultrasound experience]) was given the patient's medical record number, date of last presurgical ultrasound, nodule location within the thyroid, and tridimensional size to facilitate the identification of the biopsied nodule. When available, a dedicated presurgical lateral neck ultrasound was also reviewed by the observers. All observers used the Centricity™ Universal Viewer PACS system (GE Healthcare) to review still images and CINE clips when available, mostly from studies performed in recent years at our institution. This study was approved by the Institutional Review Board and a waiver of consent was granted.

Ultrasound evaluation

All observers were blinded to the histological outcomes and independently assessed each of the following characteristics: echogenicity of the solid portion of the nodule with respect to the normal thyroid parenchyma (hypoechoic, isoechoic, hyperechoic, or heterogeneous echogenicity –heteroechoic–); composition (solid/predominantly solid –<20% cystic component–, mixed –>20% cystic component–, spongiform, or cystic); margins (regular or irregular/microlobulated/infiltrative); echogenic foci (none/comet tails, low-risk calcifications [macrocalcifications or rim calcifications]), or high-risk calcifications (microcalcifications or interrupted rim calcifications); extrathyroidal extension (defined as bulging, abutment or disruption of the

thyroid capsule; present or absent); and suspicious lymph nodes (present or absent). A taller than wide shape in the transverse view (present or absent) was derived from the tridimensional size given to locate the nodule, thus common to all observers, and considered present when the difference between the anterior-posterior measurement exceeded (≥2 mm) the medial-lateral measurement to allow for small artefactual differences.

Nodules were classified according to the 2015 ATA guidelines and classified into one of five suspicion patterns: benign (pure cysts); very-low (spongiform or mixed nodules without eccentric solid components), low (isoechoic or hyperechoic nodules that were solid or partly cystic with eccentric solid components), intermediate (hypoechoic solid nodules), and high (hypoechoic, solid or partly cystic nodules with at least one of the following suspicious sonographic criteria: taller than wide shape; microcalcifications or interrupted rim calcifications; irregular, microlobulated or infiltrative margins; extrathyroidal extension; and presence of suspicious lymph nodes). Nodules not fitting in any of these categories were classified into a sixth "non-ATA" sonographic pattern (i.e., heteroechoic nodules with or without other suspicious features; and iso or hyperechoic nodules with at least one of the suspicious features described in the high-suspicion sonographic pattern). The final characterization of each sonographic feature and pattern was determined by agreement of at least two of the observers. In the absence of agreement, the feature was characterized by consensus of at least two of the observers during a joint review.

For the risk-stratification of non-ATA and high-suspicion sonographic patterns, nodules were separated according to the number of identified suspicious features, including shape taller than wide, irregular/lobulated/infiltrative margins, microcalcifications or interrupted rim calcifications, extrathyroidal extension, and suspicious lymph nodes. Suspicious lymph nodes were considered absent (not present) in cases without a dedicated lateral neck ultrasound available for review. Nodules in the non-ATA pattern that were iso or hyperechoic were risk-stratified separately from those that were heteroechoic.

Pathology evaluation

Cytological diagnoses were rendered using the 2009 Bethesda System for Reporting Thyroid Cytopathology by board certified cytologists and retrieved from the cytology reports [15]. Most histological diagnoses (97%) were also issued by board certified pathologists. The histology of most malignancies (87%), were re-reviewed blinded to the ultrasound characteristics. When appropriate, Follicular Variant of Papillary Thyroid Carcinomas (FVPTC's) were reclassified as Non-Invasive Follicular Thyroid Neoplasms with Papillary-like Nuclear Features (NIFTP's), or as Conventional Variant of Papillary Thyroid Carcinomas (CVPTC's). Ultrasound images of the biopsied nodule were matched to the histological diagnosis by size and location; and cases with unclear correlation were excluded.

Statistical analysis

For each individual sonographic feature and for the sonographic patterns the percent of overall agreement and Randolph's free-marginal multirater *kappa* for ($m = 3$) raters was calculated. We computed 95% confidence intervals for Kappa by non-parametric bootstrap using a first order normal approximation [16, 17]. The inter-observer agreement is considered poor for Kappa values from 0.0 (no agreement) to 0.2; fair from 0.2 to 0.4; moderate from 0.4 to 0.6; good from 0.6 to 0.8; and very good from 0.8 to 1.0 (full agreement).

Statistical Analysis was performed using SAS (version 9.4; SAS Institute, Cary, NC) and R (ver 3.4.3; R Foundation for Statistics Computing). Fisher Exact Tests were used for categorical variables, whereas Chi-square tests evaluated comparisons. 95% Confidence Intervals (CI) were used for odds ratios calculated from contingency tables. *P*-values were two-sided and considered statistically significant below 0.05.

Results

Study cohort

Out of 861 ITNs evaluated during the study period, 529 had known histological correlation. Presurgical thyroid ultrasound images were available for review in 463 (88%) nodules (176 AUS/FLUS, 38%; and 287 FN, 62%) in 415 patients (76% women, mean age 53 years), which were included in the study. Out of the patients with known thyroid function ($n = 375$, 81%) at the time of the biopsy, 5% had hyperthyroidism (defined as TSH below the reference range or patient under antithyroid drugs), 18% hypothyroidism (defined as TSH above the reference range or patient on levothyroxine), and 77% had normal thyroid function (TSH within reference range and patient not on thyroid drugs).

Sonographic features, sonographic patterns, and inter-observer agreement

Individual and consensus categorization of each sonographic features are presented in Table 1. Most nodules were isoechoic (42%), solid (89%), without echogenic foci (87%), had well-defined margins (87%), were intrathyroidal

Table 1 Frequency of sonographic features by observer

	Observer #1	Observer #2	Observer #3	Overall	*P*-value
Echogenicity					<.0001
Hypoechoic	112 (24)	80 (17)	204 (44)	117 (25)	
Isoechoic	233 (50)	97 (21)	149 (32)	196 (42)	
Hyperechoic	12 (3)	20 (4)	15 (3)	13 (3)	
Heteroechoic	106 (23)	266 (57)	95 (21)	137 (30)	
Composition					<.0001
Cystic	0	0	1 (0)	1 (0)	
Solid (or mostly solid)	395 (85)	300 (65)	432 (93)	411 (89)	
Mixed	47 (10)	157 (34)	12 (3)	40 (9)	
Spongiform	21 (5)	6 (1)	18 (4)	11 (2)	
Margin					<.0001
Regular	346 (75)	341 (74)	452 (98)	404 (87)	
Irregular/microlobulated	117 (25)	112 (24)	11 (2)	59 (13)	
Infiltrative	0	10 (2)	0	0	
Echogenic foci					<.0001
None	364 (79)	368 (79)	421 (91)	402 (87)	
Microcalcifications	20 (4)	52[a] (11)	11 (2)	17 (4)	
Macrocalcifications	37 (8)	14 (3)	8 (2)	17 (4)	
Micro and macro	13 (3)	15 (3)	12 (3)	16 (3)	
Rim (interrupted or not)	6 (1)	4 (1)	8 (2)	5 (1)	
Comet tails	23 (5)	11[a] (2)	3 (1)	6 (1)	
ETE	6 (1)	56 (12)	1 (0)	2 (0)	<.0001
Suspicious lymph nodes ($n = 116$)	12 (10)	21 (18)	16 (14)	12 (10)	0.23
ATA Sonographic pattern					0.32[b] (0.26,0.38)
Very-Low	41 (9)	8(2)	22 (5)	25 (5)	
Low	171 (37)	89 (19)	135 (29)	159 (34)	
Intermediate	52 (11)	50 (11)	159 (34)	74 (16)	
High	57 (12)	28 (6)	37 (8)	36 (8)	
Non-ATA pattern	142(31)	288 (62)	110 (24)	169 (37)	

ATA American Thyroid Association, *ETE* extrathyroidal extension, *Non-ATA* see description in methods section

[a]One nodule with microcalcifications and comet tails (counted twice in the table)

[b]Intraclass correlation coefficient with 95% confidence interval. There were no nodules classified as Benign sonographic pattern

(>99%), and did not have suspicious lymph nodes in a dedicated lateral neck ultrasound (90%).The inter-observer agreement was fair for the echogenicity; moderate for margins and composition; good for echogenic foci; and very good for extrathyroidal extension and lymph node metastasis (Table 2).

The sonographic pattern was characterized as very-low suspicion in 5%, low suspicion in 34%, intermediate suspicion in 16%, high suspicion in 8%, and non-ATA in 37% of the nodules (Table 1). The distribution of the categories of each sonographic feature was significantly different between the observers ($P < 0.0001$ in all features). Sonographic patterns between the observers had a poor intraclass correlation coefficient for absolute agreement ($r = 0.32$; 95% confidence interval 0.26–0.38). The "true" sonographic patterns determined by consensus of at least 2 of the observers; or derived from the consensus of each individual feature had very high ($r = 0.91$; $P < 0.0001$) Spearman rank correlation (Supplemental Table S1). The inter-observer agreement was fair for the ATA sonographic patterns (Table 2).

Table 2 Inter-observer agreement for sonographic features (Free-marginal kappa)

Sonographic feature (n categories)	% overall agreement	Free-marginal kappa
Echogenicity (4)	51%	0.35 (0.31–0.39)
Composition (4)	75%	0.66 (0.62–0.70)
Margins (2)	75%	0.50 (0.43–0.55)
Echogenic foci (3)	85%	0.77 (0.73–0.81)
ETE (2)	91%	0.82 (0.78–0.86)
Suspicious lymph nodes ($n = 116$) (2)	90%	0.80 (0.72–0.89)
Sonographic pattern (5)	49%	0.36 (0.32–0.41)

The following categories were used to calculate the inter-observer agreement: Echogenicity: hypoecoic, isoechoic, hyperechoic, or heteroechoic; Composition solid/predominantly solid, mixed, spongiform, or cystic; Margins: regular or irregular/microlobulated/infiltrative; Echogenic foci: none/comet tails, low-risk calcifications [macrocalcifications or rim calcifications], or high-risk calcifications [microcalcifications or interrupted rim calcifications]; ETE: present or absent; suspicious lymph nodes: present or absent; Sonographic pattern: Very-Low, Low, Intermediate, High, and non-ATA (no nodules were characterized as benign sonographic pattern, thus this category was excluded from the assessment)

ETE extrathyroidal extension

Table 3 Interpretation of sonographic features with complete disagreement during individual image review

	Observer #1	Observer #2	Observer #3	Consensus
Echogenicity ($n = 59$)				
Hypoechoic	2 (3)	3 (5)	48 (81)	1 (2)
Isoechoic	51 (86)	1 (2)	7 (12)	43 (73)
Hyperechoic	2 (3)	2 (3)	2 (3)	1 (2)
Heteroechoic	4 (7)	53 (90)	2 (3)	14 (24)
Composition ($n = 12$)				
Cystic	0 (0)	0 (0)	1 (8)	1 (8)
Solid/predominantly solid	1 (8)	1 (8)	10 (83)	8 (67)
Mixed	1 (8)	10 (83)	0 (0)	2 (17)
Spongiform	10 (83)	1 (8)	1 (8)	1 (8)
Echogenic foci ($n = 9$)				
None/comet tails	1 (11)	2 (22)	6 (67)	3 (33)
Low-risk (macrocalcifications or rim calcifications)	7 (78)	1 (11)	0 (0)	1 (11)
High-risk (microcalcifications or interrupted rim calcifications)	1 (11)	6 (67)	3 (33)	5 (56)
Sonographic pattern ($n = 76$)				
Very-low	19 (25)	1 (1)	2 (3)	6 (8)
Low	39 (51)	1 (1)	16 (21)	39 (51)
Intermediate	1 (1)	1 (1)	54 (71)	5 (7)
High	14 (18)	5 (7)	3 (4)	9 (12)
Non-ATA	3 (4)	68 (89)	1 (1)	17 (22)
Sonographic pattern only discordant feature ($n = 28$)				
Very-low	8 (29)	0 (0)	1 (4)	4 (14)
Low	6 (21)	0 (0)	7 (25)	6 (21)
Intermediate	0 (0)	0 (0)	19 (68)	4 (14)
High	13 (46)	4 (14)	1 (4)	9 (32)
Non-ATA	1 (4)	24 (86)	0 (0)	5 (18)

ATA American Thyroid Association, *Non-ATA* see description in methods section

Sources of discordance

A total of 104 thyroid nodules (22%) required joint image review due to complete disagreement in at least one of the features (Table 3, Fig. 1). The cytological diagnosis was AUS/FLUS in 54 (52%); and FN in 50 (48%) nodules. The sonographic pattern of these 104 nodules was finally classified as very-low suspicion in 13 (13%), low suspicion in 43 (41%), intermediate suspicion in 5 (5%), high suspicion in 13 (13%), and non-ATA in 30 (29%). Echogenicity interpretation was the major source of discordance. Of the 463 thyroid nodules evaluated in the study, there was complete disagreement in echogenicity in 59 (13%); composition in 12 (3%); echogenic foci in 9 (2%); and sonographic pattern in 76 (16%). Among those 76 nodules with complete disagreement in the sonographic pattern, there was agreement by at least two observers in all the other sonographic features in 28 nodules; whereas in the other 48 there was disagreement in at least one other feature, which drove the disagreement in the sonographic pattern (echogenicity in 41; composition in 5; and composition and echogenicity in 2).

Score-based risk stratification of non-ATA and high-suspicion sonographic patterns

Based on consensus (agreement ≥ 2 observers) of the sonographic pattern, 169 nodules (37%) were classified into the non-ATA pattern (Table 4). The majority of these nodules were heteroechoic ($n = 136$, 80%); and most heteroechoic nodules did not have other suspicious features. Considering NIFTPs malignant, the overall prevalence of malignancy in this cohort was 36%; 34% for nodules with 0–1 suspicious features and 53% for nodules with 2–4 features (OR 2.18 [0.75, 6.33]; $P = 0.17$). If NIFTPs were considered benign, the prevalence of malignancy would drop to 18% in nodules with 0–1 suspicious features but would remain at 53% in nodules with ≥2 suspicious features

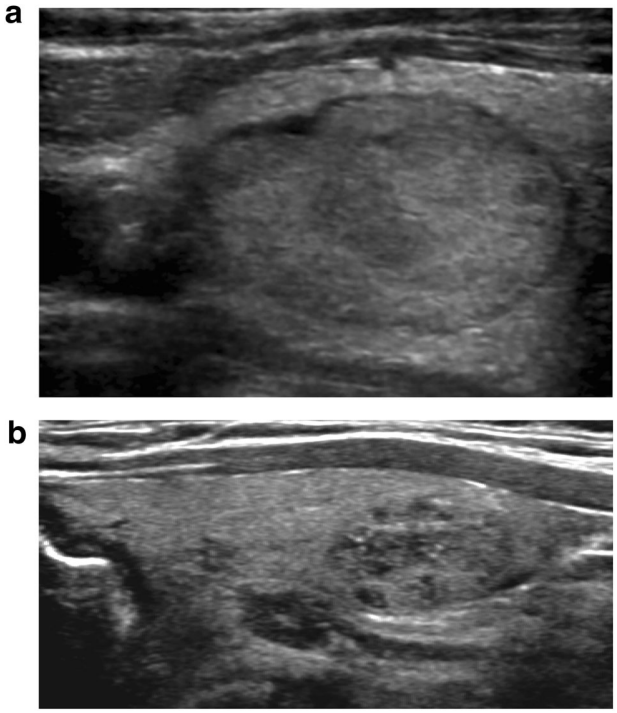

Fig. 1 Examples of thyroid nodules with complete sonographic feature disagreement. **a** Thyroid nodule with complete disagreement on echogenicity: isoechoic by observer 1, heteroechoic by observer 2, and hypoechoic by observer 3. **b** Thyroid nodule with complete disagreement on composition: spongiform by observer 1, mixed by observer 2, and predominantly solid by observer 3

(OR 5.14 [1.72, 15.4]; $P = 0.004$). Considering NIFTPs malignant, the risk of malignancy of heteroechoic nodules was higher than for hypoechoic nodules in the absence of other suspicious features (OR 2.47 [1.17, 5.20]; $P = 0.02$).

Based on consensus (agreement ≥2 observers) of the sonographic pattern, 36 nodules (8%) were classified into the high-suspicion sonographic pattern (Table 5). All nodules were hypoechoic and solid/predominantly solid. The rate of malignancy increased with the number of suspicious features: considering NIFTPs malignant 0, 35, 91, 100, and 100%; and considering NIFTPs benign 0, 30, 73, 100, and 100%, for nodules with 0, 1, 2, 3, or 4 additional suspicious features, respectively. Only three nodules met NIFTP diagnostic criteria in this group. All three NIFTPs had irregular/microlobulated margins; and two had microcalcifications as well.

Discussion

This study evaluated the inter-observer agreement for sonographic features in a large cohort of ITNs. The inter-observer agreement for echogenicity was fair and it was the greatest source of disagreement. Heteroechogenicity was associated with an increased rate of malignancy but its interpretation seems equally subjective. Furthermore, we found that the number of suspicious features in the non-ATA and high-suspicion sonographic patterns was associated with the rate of malignancy. Additionally, most (93%) NIFTPs had <2 suspicious features, and none had 3 or more suspicious features.

The inter-observer agreement for all sonographic features was moderate to good except for echogenicity, which was fair. In previous publications, echogenicity has also been among the features with lowest inter-observer agreement with kappa values around 0.35–0.55 [11–13, 18–21]. In this study, inter-observer agreement for echogenicity was 0.35 and it was the main source of disagreement. Two factors seem to have driven this low agreement: the addition of heteroechogenicity to the characterization of echogenicity; and different thresholds for the interpretation of echogenicity.

First, echogenicity has been typically characterized into three different categories: iso, hyper, and hypoechogenicity. In this study, however, a fourth category of heteroechogenicity was included. This extra category, often not characterized or analyzed as a different variable (homogeneous or heterogeneous echotexture), overlaps with all the other three categories and thus contributed significantly to a decrease in the agreement. Some classifications such as the EU-TIRADS suggest considering the most suspicious echogenicity in heterogeneous nodules [22]. It seems, however, that in the absence of other suspicious sonographic features, there is no significant difference between iso or hypoechoic cytologically indeterminate thyroid nodules [14, 23, 24]. Conversely, heteroechoic nodules had a significantly higher risk of malignancy than hypoechoic nodules without other suspicious features in our series [14]. Supporting this finding, a previous study using computer assisted diagnosis, found that a higher heterogeneity index was associated with the rate of malignancy whereas there was great overlap of mean echogenicity between iso-hypoechoic nodules, isoechoic, hypoechoic, and heteroechoic nodules [25]. Thus heteroechogenicity arises as an important sonographic feature, needing to be better defined, characterized and reported. It remains to be determined whether reporting heterogeneity as an independent variable rather than as a category of echogenicity provides diagnostic advantage. In this study, all heteroechoic nodules were classified into the non-ATA category, which drove a significantly higher rate of non-ATA pattern nodules than in previous publications [24, 26]. Interestingly, the rate of malignancy in this category was stratified by the number of additional suspicious features, which was also observed in a recent publication [26]. All NIFTPs had <2 suspicious features; whereas the prevalence of malignancy of non-ATA pattern ITNs with ≥2 suspicious features was similar to that of ITNs with high-suspicion sonographic pattern [14].

Table 4 Risk of malignancy of ITNs with non-ATA sonographic pattern according to number of suspicious features

Iso/hyper ($n = 33$)	*N* (%)	Cancer (NIFTPs malignant)	Cancer (NIFTPs benign)
0 features	3 (2)	0	0
1 feature	25 (15)	9 (36)	5 (20)
Shape T > W	12	4	3
Margin	8	4	2
Microcalcifications	5	1	0
2 features	4 (2)	3 (75)	3 (75)
Shape T > W + margin	1	1	1
Shape T > W + microcalcifications	1	0	0
Margin + microcalcifications	2	2	2
3 features	1 (1)	1 (100)	1 (100)
Shape T > W + margin + microcalcifications	1	1	1

Heteroechoic ($n = 136$)	*N* (%)	Cancer (NIFTPs malignant)	Cancer (NIFTPs benign)
0 features	87 (51)	30 (34)	16 (18)
1 feature	39 (23)	14 (36)	7 (18)
Shape T > W	12	4	0
Margin	14	7	6
Microcalcifications	8	2	1
SLN	5	1	0
2 features	5 (3)	1 (20)	1 (20)
Shape T > W + margin	1	0	0
Shape T > W + microcalcifications	1	0	0
Margin + microcalcifications	3	1	1
3 features	4 (2)	2 (50)	2 (50)
Shape T > W + margin + microcalcifications	2	0	0
Shape T > W + microcalcifications + SLN	2	2	2
4 features	1 (1)	1 (100)	1 (100)
Shape T > W + margin + microcalcifications + ETE	1	1	1

Features considered suspicious: shape taller than wide, irregular/lobulated/infiltrative margins, microcalcifications, extrathyroidal extension, and suspicious lymph nodes

SLN were considered absent (not present) in cases with no neck ultrasound available for review

Note that the number of heteroechoic nodules does not match Table 1. Three nodules were considered to be heteroechoic (agreement by at least 2 observers) during initial independent review but considered very-low suspicion pattern. Two due to spongiform composition, and one was considered to be a pseudonodule during consensus review of the sonographic pattern. Conversely, three nodules were considered hypoechoic (agreement by at least 2 observers) during initial independent review. However, there was disagreement on sonographic pattern, so the images were reviewed jointly by all three observers, who agreed that two would be better classified as heteroechoic (one of them with irregular margins), and the other one as isoechoic with irregular margins, thus all three were classified as non-ATA pattern

ETE extrathyroidal extension, *Shape T > W* shape taller than wide, *SLN* suspicious lymph nodes

Second, observers seem to have a different threshold for the interpretation of echogenicity. Most nodules were isoechoic for observer #1 (50%), heteroechoic for observer #2 (57%), and hypoechoic for observer #3 (44%). Interestingly, we found that many of the 59 nodules with complete disagreement in echogenicity characterization were borderline between iso- and slightly hypoechoic (Fig. 1) and often had areas of different echogenicity, a pattern which was characterized separately in a previous publication as iso-hypoechoic [25]. More than 80% of these nodules were characterized as isoechoic by observer #1, heteroechoic by observer #2, and hypoechoic by observer #3; suggesting

Table 5 Risk of malignancy of ITNs with high-suspicion sonographic pattern according to number of suspicious features

High-suspicion pattern ($n = 36$)	*N* (%)	Cancer (NIFTPs malignant)	Cancer (NIFTPs benign)
0 features	2 (6)	0 (0)	0 (0)
1 feature	20 (56)	7 (35)	6 (30)
Shape T > W	5	0	0
Margin	10	4	3
Microcalcifications	5	3	3
2 features	11 (31)	10 (91)	8 (73)
Shape T > W + margin	2	2	2
Margin + microcalcifications	7	6	4
Margin + ETE	1	1	1
Microcalcifications + SLN	1	1	1
3 features	1 (3)	1 (100)	1 (100)
Shape T > W + margin + microcalcifications	1	1	1
4 features	2 (6)	2 (100)	2 (100)
Shape T > W + margin + microcalcifications + ETE	1	1	1
Shape T > W + margin + microcalcifications + SLN	1	1	1

Features considered suspicious: shape taller than wide, irregular/lobulated/infiltrative margins, microcalcifications, extrathyroidal extension, and suspicious lymph nodes

SLN were considered absent (not present) in cases with no neck ultrasound available for review

All nodules were hypoechoic and solid/predominantly solid. Two nodules had no additional suspicious features despite being considered high-suspicion sonographic pattern by at least two observers because all other individual suspicious features were considered absent by at least two observers (i.e., at least two observers identified additional suspicious features but did not agree in which of them was present)

ETE extrathyroidal extension, *Shape T > W* shape taller than wide, *SLN* suspicious lymph nodes

that there was internal consistency in the interpretation. During consensus review, most of these nodules (73%) were reclassified as isoechoic. Although we cannot assess intra-observer agreement in this study, previous publications have found a significantly higher intra-observer than inter-observer agreement for echogenicity [12, 18, 27].

Echogenicity is the single most influential feature of the 2015 ATA sonographic pattern classification; as it determines whether a solid nodule is classified in the low, intermediate, or non-ATA sonographic patterns. Not surprisingly, the inter-observer agreement for the ATA sonographic patterns observed in this study was fair, driven by the poor agreement for echogenicity. A recent study also analyzed the reproducibility of sonographic features and ATA sonographic patterns for cytologically indeterminate thyroid nodules and found similar results [28]. A previous study in this same cohort did not find significant differences in the rate of malignancy between low or intermediate suspicion sonographic patterns, but it was significantly higher for nodules in the non-ATA pattern [14]. Although the characterization of echogenicity may not have an effect on patient management following an indeterminate cytology result; different sonographic patterns are associated with different size-thresholds for biopsy [7]. Thus echogenicity interpretation may lead to significant differences in the proportion of thyroid nodules selected for biopsy, which is of significant concern given the high prevalence of thyroid nodules in the general population [29]. Objective echogenic nodule interpretation seems the next necessary step to limit the number of thyroid nodules that are unnecessarily biopsied, potentially triggering additional expensive tests and diagnostic surgeries [25, 30].

Because of the retrospective nature of the study, some information was missing, mostly from patients seeking second opinion or surgical treatment at our academic cancer center. In this regard, presurgical thyroid function tests could not be found in 19% of the study cohort. Although we cannot rule out autonomy in these nodules it is very unlikely because TSH measurement was the initial step in the evaluation of thyroid nodules throughout the period of the study, the sonographic pattern was no different from that of patients in which thyroid function was known (data not shown), and the malignancy rate was higher in this subset than in the rest of the cohort either considering NIFTP malignant (44% versus 25%) or benign (33% versus 15%), a finding not expected if there were many autonomous nodules which are associated with <5% of malignancy. This study could suffer from selection bias because only nodules with histological confirmation were included, and because

lateral neck compartments were only evaluated in patients with a dedicated lateral neck ultrasound thus potentially enriching the rates of nodules with suspicious sonographic pattern and suspicious lymph nodes. Also because of the retrospective design, observers were able to review archived static images only and in some, but not all, nodule CINE images. The numbers of images available per case, and system, settings, and operator during image acquisition, were variable. As a result, image quality was variable as well, limiting the comparability of sonographic features between images. However, this is a real-world limitation as ultrasound systems and operators are different in each clinic, and settings need to be adjusted to the individual acoustic window and nodule location. Even though each observer had a different training and experience background, an interpretation consensus guide was not standardized prior to observer scoring in order to simulate a real-world scenario. Ultrasound images were reviewed independently by each of the three observers. It is possible that different screen resolutions and light environments could have had an impact on the characterization of sonographic features, particularly echogenicity. It seems, however, that there was a true difference in the interpretation of sonographic features between the three observers. The frequency of the categories within each variable was significantly different between observers.

Conclusion

This study shows that a scoring-based classification system improves risk stratification of pattern-based classifications at least for ITNs, as the number of suspicious features in non-ATA and high-suspicion patterns are directly associated with cancer risk and cancer aggressiveness. Furthermore, heterogeneity arises as a suspicious sonographic feature, at least for ITNs, thus worth reporting. The low reproducibility of echogenicity and heterogeneity found in the study, stresses the need for future studies focusing on finding an objective way to measuring them due to the significant impact these features have on the rate of malignancy and clinical management of patients with thyroid nodules.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent The study was carried with a waiver of informed consent from our Institutional Review Board.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. H.S. Ahn, H.J. Kim, H.G. Welch, Korea's thyroid-cancer "epidemic"-screening and overdiagnosis. N. Engl. J. Med. **371**(19), 1765–1767 (2014). https://doi.org/10.1056/NEJMp1409841
2. S. Nagar, B. Aschebrook-Kilfoy, E.L. Kaplan, P. Angelos, R.H. Grogan, Age of diagnosing physician impacts the incidence of thyroid cancer in a population. Cancer Causes Control **25**(12), 1627–1634 (2014). https://doi.org/10.1007/s10552-014-0467-2
3. R. Udelsman, Y. Zhang, The epidemic of thyroid cancer in the United States: the role of endocrinologists and ultrasounds. Thyroid **24**(3), 472–479 (2014). https://doi.org/10.1089/thy.2013.0257
4. S. Vaccarella, S. Franceschi, F. Bray, C.P. Wild, M. Plummer, L. Dal Maso, Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. N. Engl. J. Med. **375**(7), 614–617 (2016). https://doi.org/10.1056/NEJMp1604412
5. J.P. Zevallos, C.M. Hartman, J.R. Kramer, E.M. Sturgis, E.Y. Chiao, Increased thyroid cancer incidence corresponds to increased use of thyroid ultrasound and fine-needle aspiration: a study of the Veterans Affairs health care system. Cancer **121**(5), 741–746 (2015). https://doi.org/10.1002/cncr.29122
6. J.A. Sosa, J.W. Hanna, K.A. Robinson, R.B. Lanman, Increases in thyroid nodule fine-needle aspirations, operations, and diagnoses of thyroid cancer in the United States. Surgery **154**(6), 1420–1426 (2013). https://doi.org/10.1016/j.surg.2013.07.006. discussion 1426–1427
7. B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, F. Pacini, G.W. Randolph, A.M. Sawka, M. Schlumberger, K.G. Schuff, S.I. Sherman, J.A. Sosa, D.L. Steward, R.M. Tuttle, L. Wartofsky, 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 **26**(1), 1–133 (2016). https://doi.org/10.1089/thy.2015.0020
8. P. Campanella, F. Ianni, C.A. Rota, S.M. Corsello, A. Pontecorvi, Quantification of cancer risk of each clinical and ultrasonographic suspicious feature of thyroid nodules: a systematic review and meta-analysis. Eur. J. Endocrinol. / Eur. Fed. Endocr. Soc. **170**(5), R203–R211 (2014). https://doi.org/10.1530/eje-13-0995
9. L.R. Remonti, C.K. Kramer, C.B. Leitao, L.C. Pinto, J.L. Gross, Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. Thyroid **25**(5), 538–550 (2015). https://doi.org/10.1089/thy.2014.0353
10. J.P. Brito, M.R. Gionfriddo, A. Al Nofal, K.R. Boehmer, A.L. Leppin, C. Reading, M. Callstrom, T.A. Elraiyah, L.J. Prokop, M. N. Stan, M.H. Murad, J.C. Morris, V.M. Montori, The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. J. Clin. Endocrinol. Metab. **99**(4), 1253–1263 (2014). https://doi.org/10.1210/jc.2013-2928
11. W.J. Moon, S.L. Jung, J.H. Lee, D.G. Na, J.H. Baek, Y.H. Lee, J. Kim, H.S. Kim, J.S. Byun, D.H. Lee, Benign and malignant thyroid nodules: US differentiation-multicenter retrospective study. Radiology **247**(3), 762–770 (2008). https://doi.org/10.1148/radiol.2473070944
12. C.S. Park, S.H. Kim, S.L. Jung, B.J. Kang, J.Y. Kim, J.J. Choi, M. S. Sung, H.W. Yim, S.H. Jeong, Observer variability in the

sonographic evaluation of thyroid nodules. J. Clin. Ultrasound **38** (6), 287–293 (2010). https://doi.org/10.1002/jcu.20689
13. P. Valderrabano, D.L. Klippenstein, J.B. Tourtelot, Z. Ma, Z.J. Thompson, H.S. Lilienfeld, B. McIver, New American Thyroid Association Sonographic Patterns for Thyroid Nodules Perform Well in Medullary Thyroid Carcinoma: Institutional Experience, Systematic Review, and Meta-Analysis. Thyroid **26**(8), 1093–1100 (2016). https://doi.org/10.1089/thy.2016.0196
14. P. Valderrabano, M.J. McGettigan, C.A. Lam, L. Khazai, Z.J. Thompson, C.H. Chung, B.A. Centeno, B. McIver, Thyroid nodules with indeterminate cytology: utility of the American Thyroid Association Sonographic Patterns for Cancer Risk Stratification. Thyroid **28**(8), 1004–1012 (2018). https://doi.org/10.1089/thy.2018.0085
15. E.S. Cibas, S.Z. Ali, The Bethesda system for reporting thyroid cytopathology. Thyroid **19**(11), 1159–1165 (2009). https://doi.org/10.1089/thy.2009.0274
16. Randolph, J. *Free-marginal multirater kappa: an alternative to Fleiss' fixed-marginal multirater kappa*. Joensuu University Learning and Instruction Symposium, Joensuu, Finland, 2005.
17. M.J. Warrens, Inequalities between multi-rater kappas. Adv. Data Anal. Classif. **4**(4), 271–286 (2010). https://doi.org/10.1007/s11634-010-0073-4
18. S.H. Choi, E.K. Kim, J.Y. Kwak, M.J. Kim, E.J. Son, Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid **20**(2), 167–172 (2010). https://doi.org/10.1089/thy.2008.0354
19. S.H. Park, S.J. Kim, E.K. Kim, M.J. Kim, E.J. Son, J.Y. Kwak, Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules. AJR Am. J. Roentgenol. **193**(5), W416–W423 (2009). https://doi.org/10.2214/AJR.09.2541
20. G. Grani, L. Lamartina, V. Cantisani, M. Maranghi, P. Lucia, C. Durante, Interobserver agreement of various thyroid imaging reporting and data systems. Endocr. Connect **7**(1), 1–7 (2018). https://doi.org/10.1530/EC-17-0336
21. W. Phuttharak, A. Boonrod, V. Klungboonkrong, T. Witsawapaisan, Interrater Reliability of Various Thyroid Imaging Reporting and Data System (TIRADS) Classifications for Differentiating Benign from Malignant Thyroid Nodules. Asian Pac. J. Cancer Prev. **20**(4), 1283–1288 (2019). https://doi.org/10.31557/APJCP.2019.20.4.1283
22. G. Russ, S.J. Bonnema, M.F. Erdogan, C. Durante, R. Ngu, L. Leenhardt, European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur. Thyroid J. **6**(5), 225–237 (2017). https://doi.org/10.1159/000478927
23. P. Trimboli, M. Deandrea, A. Mormile, L. Ceriani, F. Garino, P. P. Limone, L. Giovanella, American Thyroid Association ultrasound system for the initial assessment of thyroid nodules: use in stratifying the risk of malignancy of indeterminate lesions. Head. Neck **40**(4), 722–727 (2018). https://doi.org/10.1002/hed.25038
24. T.G. Rocha, P.W. Rosario, A.L. Silva, M.B. Nunes, T.H. Silva, P. H.L. de Oliveira, M.R. Calsolari, Ultrasonography Classification of the American Thyroid Association for Predicting Malignancy in Thyroid Nodules >1cm with Indeterminate Cytology: A Prospective Study. Horm. Metab. Res **50**(8), 597–601 (2018). https://doi.org/10.1055/a-0655-3016
25. G. Grani, M. D'Alessandri, G. Carbotta, A. Nesca, M. Del Sordo, S. Alessandrini, C. Coccaro, R. Rendina, M. Bianchini, N. Prinzi, A. Fumarola, Grey-scale analysis improves the ultrasonographic evaluation of thyroid nodules. Med. (Baltim.) **94**(27), e1129 (2015). https://doi.org/10.1097/MD.0000000000001129
26. L. Gao, X. Xi, J. Wang, X. Yang, Y. Wang, S. Zhu, X. Lai, X. Zhang, R. Zhao, B. Zhang, Ultrasound risk evaluation of thyroid nodules that are "unspecified" in the 2015 American Thyroid Association management guidelines: a retrospective study. Med. (Baltim.) **97**(52), e13914 (2018). https://doi.org/10.1097/MD.0000000000013914
27. J.E. Lim-Dunham, I. Erdem Toslak, K. Alsabban, A. Aziz, B. Martin, G. Okur, K.C. Longo, Ultrasound risk stratification for malignancy using the 2015 American Thyroid Association Management Guidelines for Children with Thyroid Nodules and Differentiated Thyroid Cancer. Pedia. Radio. **47**(4), 429–436 (2017). https://doi.org/10.1007/s00247-017-3780-6
28. G. Grani, L. Lamartina, V. Ascoli, D. Bosco, F. Nardi, F. D'Ambrosio, A. Rubini, L. Giacomelli, M. Biffoni, S. Filetti, C. Durante, V. Cantisani, Ultrasonography scoring systems can rule out malignancy in cytologically indeterminate thyroid nodules. Endocrine **57**(2), 256–261 (2017). https://doi.org/10.1007/s12020-016-1148-6
29. G. Russ, S. Leboulleux, L. Leenhardt, L. Hegedüs, Thyroid incidentalomas: epidemiology, risk stratification with ultrasound and workup. Eur. Thyroid J. **3**(3), 154–163 (2014). https://doi.org/10.1159/000365289
30. M. Sollini, L. Cozzi, A. Chiti, M. Kirienko, Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: Where do we stand? Eur. J. Radio. **99**, 1–8 (2018). https://doi.org/10.1016/j.ejrad.2017.12.004