Evaluation of the departmental inter-rater reliability and accuracy when scoring thyroid nodules according to the BTA U-classification model. Is there significant disagreement?

Rtam N. Msc (Medical Ultrasound)

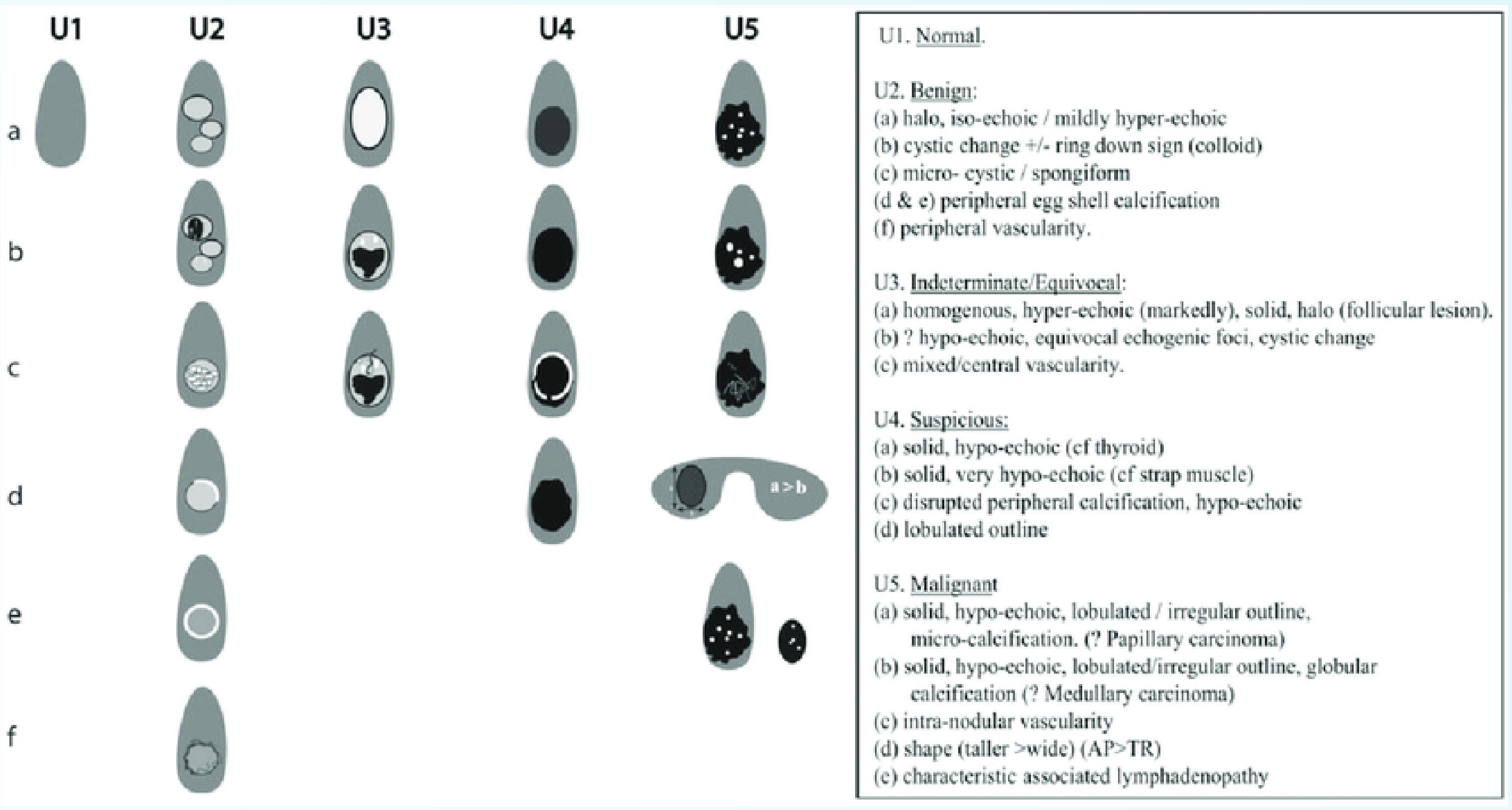
Ultrasound Department, Yeovil District Hospital

Introduction

The BTA U-classification is a risk stratification model which grades thyroid nodules (TNs) in U2-5 based on their sonographic appearance [1] . Existence of variability between the operators when U-scoring is reported in literature with some anecdotal evidence found in the author's department [2].

The aim of this study was to investigate whether there was significant disagreement when U-scoring in the department.

The objective were to assess the overall inter-operator reliability for the U categories (U2-5), for the indication of a FNAB and for each ultrasound features (echogenicity, shape, margin etc). Because a high inter-rater reliability would only indicate 'consistency' between raters and it does not demonstrate whether the operators use the model correctly, the departmental accuracy measured against the most experienced ultrasound operator in the department.



U1. Normal.

U2. Benign:
(a) halo, iso-echoic / mildly hyper-echoic
(b) cystic change +/- ring down sign (colloid)
(c) micro- cystic / spongiform
(d & e) peripheral egg shell calcification
(f) peripheral vascularity.

U3. Indeterminate/Equivocal:
(a) homogenous, hyper-echoic (markedly), solid, halo (follicular lesion).
(b) ? hypo-echoic, equivocal echogenic foci, cystic change
(c) mixed/central vascularity.

U4. Suspicious:
(a) solid, hypo-echoic (cf thyroid)
(b) solid, very hypo-echoic (cf strap muscle)
(c) disrupted peripheral calcification, hypo-echoic
(d) lobulated outline

U5. Malignant
(a) solid, hypo-echoic, lobulated / irregular outline, micro-calcification. (? Papillary carcinoma)
(b) solid, hypo-echoic, lobulated/irregular outline, globular calcification (? Medullary carcinoma)
(c) intra-nodular vascularity
(d) shape (taller >wide) (AP>TR)
(e) characteristic associated lymphadenopathy

Fig.1 BTA U-classification model. *BTA guideline, 2014 (1)*

Methodology

Eight operators, Consultant Radiologists and sonographers, were recruited to retrospectively grade 33 TNs and answer a tick box questionnaire which uses the BTA lexicon.

The inter-operator reliability was calculated using the K statistic to adjust for agreement due to chance, and in particular the Fleiss's [3] elaboration for more than two rater and variables. The confidence interval (CI), percentage agreement and Gwet AC1 agreement were also calculated. Gwet AC1[4] resolves for the paradox problem of the Kappa statistic and is indicated when there is prevalence in the population. A P value of < 0.05 was considered statistically significant.

Landis and Koch [5] was utilised for the interpretation of the reliability coefficients. K value between 0 and 0.20 corresponded to only slight agreement, between 0.20 and 0.40 to fair agreement, 0.40 and 0.60 moderate agreement, 0.60 and 0.80 substantial agreement and above 0.80 to almost perfect agreement.

In order to evaluate the departmental accuracy when U-scoring, percentage agreement (PA), Gwet AC1 and the Cohen's Kappa was calculated between the scores of each assessor and those of the most experienced operator in the department (Consultant Sonographer).

Results

U-classification:
Fair agreement (Fleiss-K=0.21) was obtained between all the participants when U-scoring (U2-5). Fair to moderate agreement was noted between sonographers (K=0.40). Significant variability was demonstrated between Radiologists (P=0.10). U5 was the most agreed upon category (K=0.56) and U3 and U4 were the least agreed on (K=0.12 and 0.19 respectively).

Recommendation for FNAB:
Indication for FNAB showed fair agreement for Radiologists' (AC1=0.34), almost substantial agreement for sonographers (AC1=0.58) and moderate overall agreement (AC1=0.41).

	Fleiss K	***Confidence Interval 95%***	***P value***
Grade			
U2	0.22	0.15 - 0.28	<0.00
U3	0.12	0.05 - 0.17	<0.00
U4	0.19	0.12 - 0.25	<0.00
U5	0.56	0.50 - 0.63	<0.00
Radio Ag.	0.13	-0.03 - 0.29	0.10
Sonog Ag.	0.40	0.22 - 0.57	<0.00
Total Ag.	0.21	0.10 - 0.32	<0.00
	Gwet AC1/PA(%)		
FNAB U2 />U2			
Total Ag.	0.41/66%	0.21 - 0.61	<0.00
Radio Ag.	0.34/60%	0.11 - 0.55	<0.00
Sonog Ag.	0.58/74%	0.35 - 0.81	<0.00

Table 1 Overall Inter-operator reliability for the U2-5 categories and indication for FNAB.

Agreement on US features:
No significant variability was measured for echogenicity (K=0.29), composition (K=0.33), shape (K=0.58), margin (K=0.45), halo (K=0.33) and vascularity (K=0.44). Significant variability was noted for the "micro-cystic/spongiform "feature that agreement due to chance could not be excluded (P>0.05).

When analysing the data for "peripheral egg calcification" and "disrupted egg calcification", the operators used these terms interchangeably which indicates necessity to revise the meaning of these terms.

Although 61 cases on 264 were reported as "taller than wider" which has high sensitivity and specificity for TN malignancy according to the BTA classification, 11 (18%) of these were classified as U2 by sonographers in 5 cases and by Radiologists in 6 cases for unclear reasons.

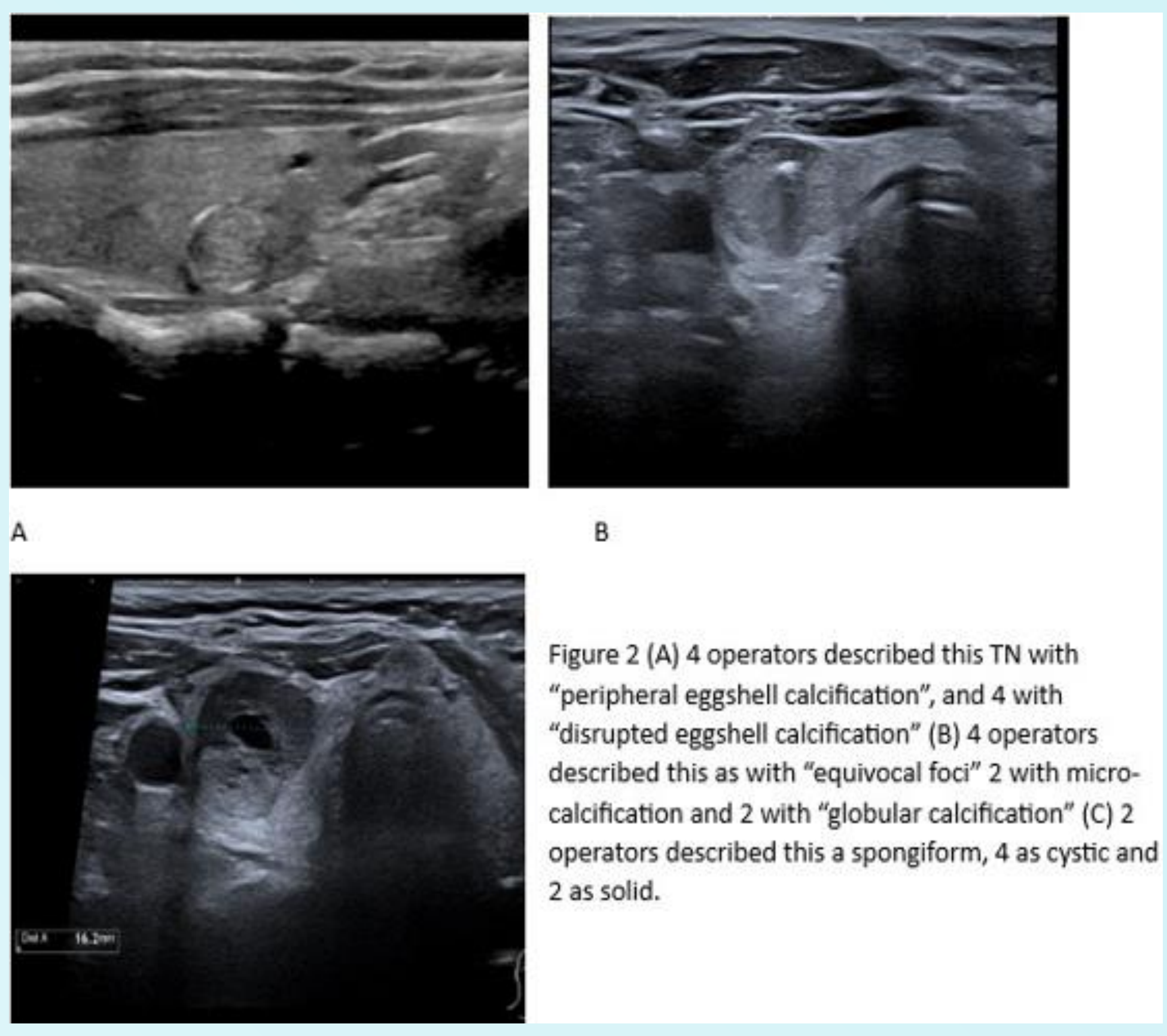


Figure 2 (A) 4 operators described this TN with "peripheral eggshell calcification", and 4 with "disrupted eggshell calcification" (B) 4 operators described this as with "equivocal foci" 2 with micro-calcification and 2 with "globular calcification" (C) 2 operators described this a spongiform, 4 as cystic and 2 as solid.

U-classification accuracy:
The participants' accuracy was measured against the expert gold standard. The overall agreements were mean Cohen's Kappa of 0.29 (Table 2). For the category U2-5, sonographers demonstrated the highest agreement with the gold standard (mean PA of 54.25% and Cohens K of 0.36). Radiologists' mean PA and Cohens K values were 44% and 0.22, however, for two Radiologists, agreement due to chance could not be excluded suggesting significant variability compared to the gold standard.

	Radiologists	**Sonographers**	**Total**
U2	56.70%	56.80%	56.75%
U3	58.35%	79.15%	68.75%
U4	18.75%	40.62%	29.70%
U5	27.87%	34.60%	31.25%
Total Ag. U2-5	44% K=0.22	54.25% K=0.36	47% K=0.29
Ag. On FNAB U2/>U2	69.7% AC1=0.46	77.27% AC1=0.60	73% AC=0.53

Table 2 Operators' accuracy against the gold standard.

Accuracy on recommendation for FNAB:
The overall accuracy for FNAB recommendation was moderate (mean of AC1 0.53). The accuracy on TNs' requirement of a FNAB for the sonographers' group was mean of PA 77.3% and AC1 0.60, while for Radiologists was PA 69.7% and AC1 0.46. Agreement due to chance could not be excluded for one Radiologist (P=0.07).

Conclusion

This study demonstrated that there is no significant inter-rater variability in U-scoring or recommending FNAB between all the US operators in the department. The study showed, however, margin for improvement particularly for the Radiologist' group (significant variability in U-scoring and lower agreement with the gold standard).

Reliability and accuracy could be improved by addressing those problematic categories and features identified with this study such as U3-U4, "micro-cystic/spongiform" vs "cystic change", "peripheral egg calcification" vs "disrupted egg calcification" and "taller than wider".

Recommendations

- To differentiate U3 nodule from U4 look at the echogenicity. If hypoechoic compared to thyroid will always be >U3.
- If there are few benign features and one suspicious feature such as "taller than wider" or "micro/macro calcification" the TN will need FNA. Favour the suspicious feature.
- Familiarise with the meaning of the terms "micro-cystic/spongiform", "cystic change", "peripheral egg calcification" vs "disrupted egg calcification" and "taller than wider".

References

1. British Thyroid Association. British Thyroid Association guidelines for the management of thyroid cancer. Clinical Endocrinology 2014: 81(1) 1-122.
2. Couzins M, Forbes S, Vigneswaran G, Mitra I, Rutherford EE. Ultrasound grading of thyroid nodules using the BTA U-scoring guidelines - Is there evidence of intra-and interobserver variability? Ultrasound. 2021 May;29(2):100-105
3. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. *1971; 76*(5): 378–382.
4. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013 Apr 29;13:61.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.