ORIGINAL RESEARCH

Image Features-based Learning Effectively Improves Inter-Observer Agreement for Beginners in Evaluating Thyroid Nodule with Ultrasound

Ying Wang, MD [a,1], Luying Gao, MD [a,1], Yuxin Jiang, MD [a], Hui Pan, MD [b], Jun Zhao, MA [b], Xin Zhou, MA [b], Qiong Wu, MM [a], Ruyu Liu, MM [a], Bo Zhang, MD [a,]*

[a] *Department of Ultrasound, Chinese Academy of Medical Sciences & Peking Union Medical College Hospital, Beijing, China;*
[b] *Department of Education, Chinese Academy of Medical Sciences & Peking Union Medical College Hospital, Beijing, China*
Received October 10, 2018; revision requested February 10, 2019; revision received November 18, 2018; accepted November 20.
[1] *Ying Wang and LuYing Gao contributed equally to this work and should be considered as co-first authors.*

Objective: Thyroid nodules are a common medical problem in China and in other parts of the world. Many guidelines use ultrasound (US) as the first choice for evaluating thyroid nodules. A major limitation of US is operator dependency, resulting in a variety of discrepancies in diagnosing thyroid nodules in the literatures. Risk stratification of thyroid nodules is based on the patterns of US features in the 2015 American Thyroid Association (ATA) management guidelines for adult patients with thyroid nodules. We hypothesize that special training targeting features recognition may increase the inter-observer agreement.

Methods: The study was conducted on 52 participants from Peking Union Medical College Hospital (PUMCH) from March to May 2018.The participants were divided into two groups by their own decision for their convenience. Image features-based learning (IF-BL) was used to train the participants to learn special features including shape, margin, echo level, internal structure, calcification, vascularity through 10 standard images based on the 2015 ATA guideline. Group A (27 subjects) received IF-BL during the first month, and Group B (25 subjects) received IF-BL during the second month. All participants evaluated US features and risk stratification in 60 US images of 20 thyroid nodules before and after the training. The test results were graded by a teaching assistant according to the rule of 0.5 points assigned to every feature and 2 points assigned to risk stratification, with a total of 100 points. Inter-observer agreements of US features and risk stratification were assessed and compared before and after the training.

Results: After the first month, Group A had better scores than Group B, the control group of the month (75.4±9.4 vs 68.7±8.4, p = 0.01). At the end of the second month during which both groups were trained, there was no difference of scores between Group A and Group B (74.5±10.4 vs 75.1±7.4, $P = 0.78$). Scores of all participants were significantly higher than the initial (74.8±9.0 vs 65.8±13.6, $P < 0.01$). After the training, the kappa values of US features improved from 0.28-0.43 to 0.43-0.75, and those of risk stratification improved from 0.13 to 0.55.

Conclusion: IF-BL can effectively help trainees correctly recognize US features and evaluate the risk stratification of thyroid nodule and can improve the inter-observer agreement.

Key words: Thyroid nodule; Cancer; ATA; Training

Advanced Ultrasound in Diagnosis and Therapy 2019;01:001–005 DOI: 10.37015/AUDT.2019.190801

** Corresponding author: Department of Ultrasound, Peking Union Medical College Hospital, 9 Dongdansantiao, Beijing 100730, China. e-mail: thyroidus@163.com*

Thyroid nodules are a common medical disease with a prevalence of 39.7–50.3% in China and in other parts of the world [1,2]. Most nodules are non-palpable, asymptomatic, and detected by ultrasound (US) [2,3]. US is useful not only for detection but also for lesions evaluation. It is the first choice for evaluating thyroid nodules in many guidelines, including the latest 2015 American Thyroid Association (ATA) management guideline [4-8]. However, the main limitation of US is its operator dependency, resulting in a variety of discrepancies in diagnosing thyroid nodules in the literatures. Moon et al. reported that there was a low degree of inter-observer agreement for thyroid nodule evaluation by US [9]. This factor limits the accuracy in applying many US guidelines. Since the risk stratification of thyroid nodule is based on the patterns of US features, it's very important to recognize US features correctly. In this study, our goal was to assess the effect of image features based-learning (IF-BL) on the inter-observer agreement using the 2015 ATA Management Guidelines as the reference.

Patients and Methods

Participants, training and tests for US image reading

This study was approved by the ethics committee of the Peking Union Medical College Hospital (PUMCH). Participants who potentially met the inclusion criteria were those with theoretical competence and limited practical experience in general US. The trainee enrollment was conducted from March to May 2018. After exclusion, 64 participants from PUMCH were invited to participate in the training program, and 52 of them completed the study. The reasons for the withdrawals of the 12 trainees included schedule conflicts, family issues, personal choice and others. The participants were divided into two groups by their own decision for their convenience, and researcher classified the subjects when the numbers of the two groups were unmatched. IF-BL was used for training according to the 2015 ATA guideline to help the participants learn special features including shape, margin, echo level, internal structure, calcification, vascularity through 10 standard images such as US images shown in Figure 1. Figure 1 presents the examples of different features of thyroid nodules for participants to learn in our training program, and all the characteristics of the nodules were typical and impressive for beginners to learn and master.

Group A (27 subjects) received IF-BL during the first month, and Group B (25 subjects) received IF-BL during the second month. Apart from receiving training in our study, all trainees had participated in traditional US classroom teaching and had had daily clinical practice in the US department, even though they are beginners to practice experience

US features and risk stratification in the 60 US images of 20 thyroid nodules were evaluated independently by all participants before and after the training. The test results were graded by a teaching assistant according to the rule of 0.5 points assigned to every feature and 2 points to risk stratification, and 100 points in total. The inter-observer agreement of US features and risk stratification before and after the training were assessed and compared.

Statistical analysis

The quantitative data are presented as means± standard deviation (SD). The qualitative data are presented as frequencies. The Shapiro-Wilk test was used to determine whether the data were normally distributed. For the parametric data, the unpaired t test was used to evaluate the differences between the two groups. For the nonparametric data, the differences between the groups were analyzed and the Mann-Whitney U test was used for this analysis. The chi-square test with Yates' correction and Fisher's exact test were used to compare categorical variables. The statistical analyses were performed with the SPSS (Version 19.0, SPSS Chicago, IL, USA) software.

To assess the inter-observer variability, the kappa value was calculated, and the kappa static was used to determine the reproducibility of the US assessment of thyroid nodule risk stratification and their associated features. A kappa value of > 0.80 was considered "excellent" or "almost perfect"; a kappa value of > 0.60-0.80 indicated "substantial"; a kappa value of > 0.40-0.60 was considered "moderate"; a kappa value of > 0.20-0.40 was fair; and a kappa value of ⩽ 0.20 indicated "poor" performance. Values of $p < 0.05$ were considered as significant. All statistical analyses were performed with the SPSS (Version 19.0, SPSS Chicago, IL, USA) software and the Stata version 10 software (StataCorp, Lake way Drive College Station, TX, USA).

Results

Baseline characteristics

The differences between the baseline characteristics of the trainees in Group A (27 subjects) and those in Group B (25 subjects) are presented in Table 1. The baseline characteristics of the two groups were comparable with respect to gender, age, education, majors in medical imaging, interest in studying US, family members suffering from thyroid disease and experience in thyroid ultrasound practice. All parameters were similar in both groups ($P > 0.05$) (Table 1).

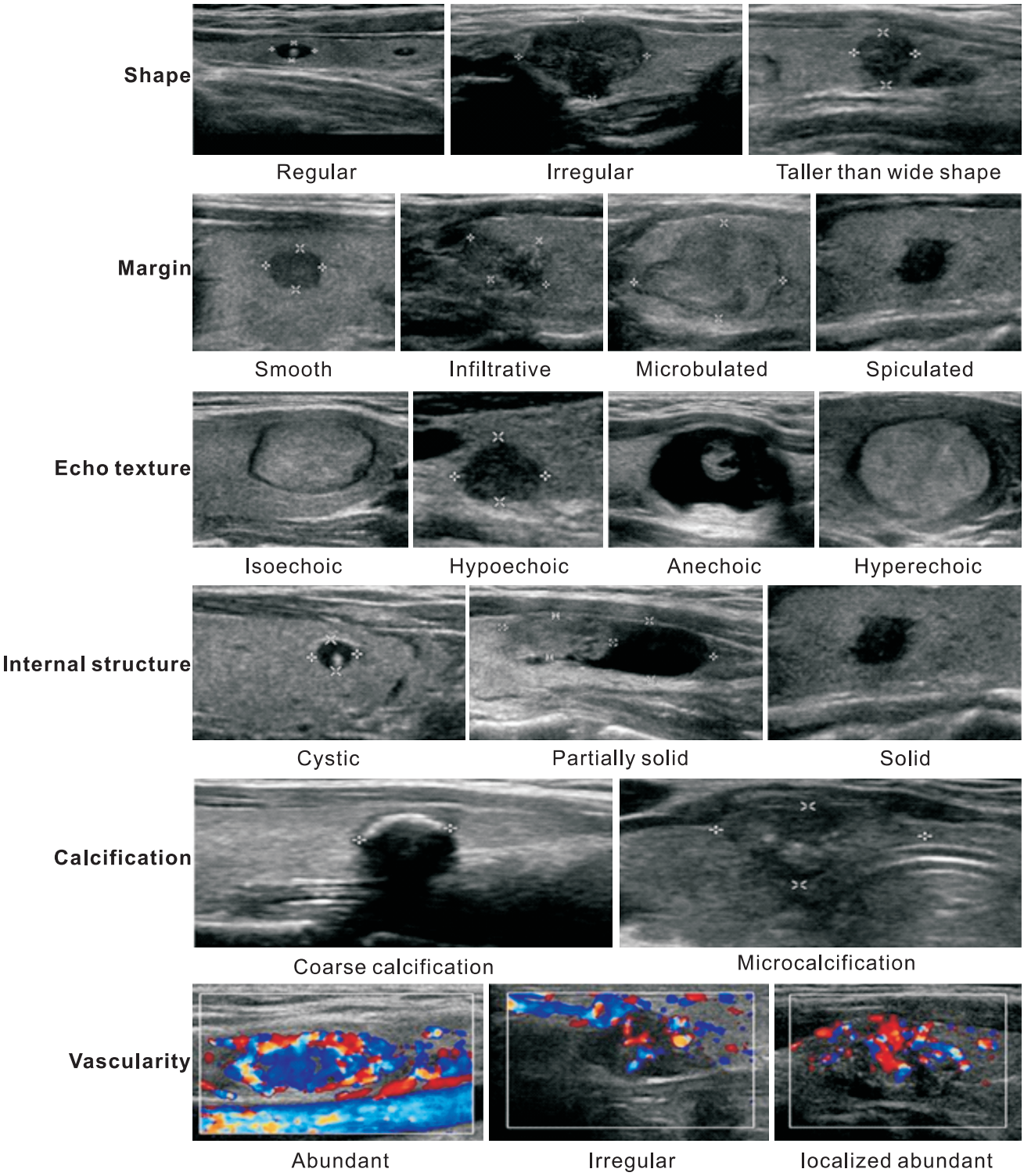

Figure 1 US features of thyroid nodule

Before entering the training program, all participants underwent a baseline US imaging test. No significant difference was found on the test scores between Group A and Group B (65.6±16.1 vs 66.1±10.4, P = 0.91) (Table 1).

Evaluation of test scores of Group A and of all participants after IF-BL

After the first month of IF-BL training, participants in Group A achieved much higher test scores than those in control group (Group B) (75.4±9.4 vs 68.7±8.4, P = 0.01).

Table 1 The baseline characteristics of Group A and Group B

Item	Group A (n = 27)	Group B (n = 25)	P value
Sex (male/female)	4/23	3/22	0.545
Age, yr (mean ± SD)	30.8±10.1	28.4±4.4	0.110
Education years, yr (mean ± SD)	16.8±5.6	18.2±2.4	0.240
Majored in medical imaging (yes/no)	9/18	11/14	0.310
Interest in studying ultrasound (passionate/average)	22/5	20/5	0.580
Relatives suffering from thyroid disease (yes/no)	16/11	16/9	0.470
Experienced in thyroid ultrasound (yes/no)	4/23	8/17	0.130
Thyroid nodule image test points (100 possible points)	65.6±16.1	66.1±10.4	0.910

Group A: Fellows received training during the first month; Group B: Fellows received training during the second month

At the end of the second month, only 12 participants in Group B completed the test. After all trainees (26 participants) in Group A and Group B completed IF-BL, there was no difference of scores between Group A and Group B (74.5±10.4 vs 75.1±7.4, P = 0.78). Test scores were significantly higher after the training than before it (74.8±9.0 vs 65.8±13.6, P < 0.01).

Inter-observer agreement for US thyroid nodule evaluation before and after IF-BL

The agreement and the strength of agreement among the trainees regarding thyroid nodule US features and risk stratification are displayed in Table 2. Before the training, the kappa values of US features ranged from 0.28 to 0.43; the agreement for shape and internal construction were moderate with corresponding values of 0.42 and 0.43; the agreement for margin, echogenicity, calcification and vascularity were fair with corresponding values ranging from 0.28 to 0.33. After the IF-BL process, all kappa values significantly improved, the improvement ranging from 0.43 to 0.75. Shape, internal struction and calcification improved with values of 0.66, 0.71 and 0.75, respectively. Margin, echogenicity, vascularity and risk stratification showed moderate agreement with values ranging from 0.43 to 0.55. The Kappa value of risk stratification improved from poor to moderate, with the corresponding value raised from 0.13 to 0.55.

Table 2 Inter-observer agreement for US thyroid nodule evaluation after training

Thyroid nodule US	Before IF-BL		After IF-BL	
	kappa	Agreement strength	kappa	Agreement strength
Shape	0.42	Moderate	0.66	Substantial
Margin	0.29	Fair	0.55	Moderate
Echogenicity	0.34	Fair	0.51	Moderate
Internal struction	0.43	Moderate	0.71	Substantial
Calcifications	0.33	Fair	0.75	Substantial
Vascularity	0.28	Fair	0.43	Moderate
Risk stratification	0.13	Slight	0.55	Moderate

All values calculated using kappa statics

Discussion

US is the first choice for evaluating thyroid nodules in many guidelines, and high diagnostic accuracy and inter-observer agreement are very important. However, the limitation of US is its operator dependency. The reproducibility among US radiologists has been

examined in different studies conducted on this subject. In consistent with the results by Moon et al, a study conducted by Chang et al. found a fair to substantially acceptable inter-observer variation of thyroid nodule US features evaluation among five US radiologists [9,10]. Moreover, a study performed by Seon et al examined the interpretation and grading of thyroid nodules among four radiologists. The reproducibility was fair to substantial [11]. The results showed there were multiple degrees of inter-observer agreements in the US feature descriptions of thyroid nodules. Therefore, to improve US diagnostic accuracy and inter-observer agreement, a process of US features recognition training is needed, which has been noted by some researchers already. In 2016, a study by M. Naren et al reported a high inter-observer reproducibility of US evaluation of thyroid nodules using the TIRADS system among six US radiologists [12]. This study emphasized the importance of regular training of trainees, but the authors did not illustrate the detailed training process. Our objective was to determine the effect of IF-BL on the diagnostic accuracy and inter-observer variability for assessment of US features and risk evaluation of thyroid nodules using ATA guidelines. After the IF-BL process, the trainees exhibited improved diagnostic accuracy. Before training, a slight-to-moderate degree of inter-observer agreement was exhibited, suggesting a poor grasp of US features and risk stratification of thyroid nodules. After a short time of the IF-BL training process, inter-observer agreement of all features and the risk stratification evaluation showed a distinct increase.

Conclusion

Our study proposed a standardized and systematic training process for beginners focusing on targeting features recognition using the 2015 ATA guideline. Participants in the study were beginners with limited US practical experience. After the IF-BL training, US diagnostic accuracy and inter-observer agreement for thyroid nodule evaluation were significantly improved in a short time. The findings of this study showed that IF-BL can effectively help trainees correctly recognize US features and evaluate risk stratification of thyroid nodules. To further improve diagnostic accuracy and inter-observer agreement of US, not only for thyroid nodule evaluation, more attention should be focused on standard training among US radiologists.

Conflicts of Interest

Disclosures: None – all authors claim no conflicts of interest or disclosures.

Funding: This study was supported by the National Natural Science Foundation of China (61672077), the Peking Union Medical College Education Foundation (Grant: 2014zlgc0136 and 2016zlgc0108).

Ethical approval: This study was approved by Peking Union Medical College Hospital's Institutional Review Board on 2016-12-1(S-K186)

References

[1] Guo H, Sun M, He W, Chen H, Li W, Tang J, et al. The prevalence of thyroid nodules and its relationship with metabolic parameters in a Chinese community-based population aged over 40 years. *Endocrine* 2014;45:230-5.

[2] Hegedüs L. Clinical practice. The thyroid nodule. *N Engl J Med* 2004;351:1764-71.

[3] Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Ann Intern Med* 1997;126:226–31.

[4] Leenhardt L, Erdogan MF, Hegedus L, Mandel SJ, Paschke R, Rago T, et al. 2013 European thyroid association guidelines for cervical ultrasound scan and ultrasound-guided techniques in the postoperative management of patients with thyroid cancer. *Eur Thyroid J* 2013;2:147-59.

[5] Perros P, Boelaert K, Colley S, Evans C, Evans RM, Gerrard Ba G, et al. Guidelines for the management of thyroid cancer. *Clin Endocrinol (Oxf)* 2014;81:1-122.

[6] Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean society of thyroid radiology consensus statement and recommendations. *Korean J Radiol* 2016;17:370-95.

[7] Gharib H, Papini E, Paschke R, Duick DS, Valcavi R, Hegedüs L, et al. American Association of Clinical Endocrinologists, Associazione Medici Endocrinologi, and European Thyroid Association Medical guidelines for clinical practice for the diagnosis and management of thyroid nodules: executive summary of recommendations. *EndocrPract* 2016;22:622-39.

[8] Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016;26:1-133.

[9] Moon WJ, Jung SL, Lee JH, Na DG, Baek JH, Lee YH, et al. Benign and malignant thyroid nodules: US differentiation--multicenter retrospective study. *Radiology* 2008;247:762–70.

[10] Park CS, Kim SH, Jung SL, Kang BJ, Kim JY, Choi JJ, et al. Observer variability in the sonographic evaluation of thyroid nodules. *J Clin Ultrasound* 2010;38:287-93.

[11] Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* 2010;20:167-72.

[12] Srinivas MN, Amogh VN, Gautam MS, Prathyusha IS, Vikram NR, Retnam MK, et al. A prospective study to evaluate the reliability of thyroid imaging reporting and data system in differentiation between benign and malignant thyroid lesions. *J Clin Image Sci* 2016;6:5.