Observational Study

Medicine

OPEN

Interobserver variability in ultrasound assessment of thyroid nodules

Jaber Alyami, PhD[a,b,c,*], Fahad F. Almutairi, PhD[a,b,c], Sultan Aldoassary, PhD[d], Amani Albeshry, BSc[d], Ali Almontashri, MD[d], Mazen Abounassif, MD[d], Majed Alamri, PhD[b]

Abstract

The first diagnostic tool for thyroid disease management is ultrasound. Despite its importance, ultrasound is an extremely subjective procedure that requires a high level of performance skill. Few studies have assessed thyroid ultrasound performance and its effectiveness, particularly the variability between observers in the assessment of ultrasound images. This study evaluated the variability in ultrasound assessments and diagnoses of thyroid nodules between 2 radiologists. In this retrospective study, 75 thyroid nodules in 39 patients were reviewed by 2 experienced radiologists. The nodule composition, margin, shape, calcification, and vasculitis were determined using echogenicity. The study evaluation included these 5 assessments and the final diagnosis. Interobserver variation was determined using Cohen kappa statistics. The interobserver agreements in the interpretation of echogenicity, shape, and margin were fair ($\kappa = 0.21–0.40$), whereas there were substantial agreements for vascularity and calcification ($\kappa = 0.62–0.78$). The agreements between the observers for individual ultrasound features in this study were the highest for vascularity and the presence/absence of calcification. The interobserver reproducibility for thyroid nodule ultrasound reporting was adequate, but the diagnostic evaluation ability of the observers was inconsistent. The variability in the interpretation of sonographic features could influence the level of suspicion of thyroid malignancy. This study emphasizes the need for consistency in the training of sonographic interpretation of thyroid nodules, particularly for echogenicity, shape, and margin.

Abbreviations: ACR = American College of Radiology, FNAC = fine-needle aspiration cytology, KSMC = King Saud Medical City, TI-RADS = thyroid imaging reporting and data system.

Keywords: agreement, interobserver variability, interpretation, reporting, thyroid nodules, ultrasound

1. Introduction

Thyroid nodules are common worldwide, and as many as 68% of adults are reported to harbor at least one nodule.[1,2] Thyroid cancer is the second-most common malignancy among women in Saudi Arabia. The increasing prevalence of thyroid nodules corresponds with the increasing number of nodules that are being revealed during cross-sectional imaging for nonthyroidal reasons; these often undergo further evaluations with a dedicated thyroid ultrasound exam and fine-needle aspiration cytology (FNAC) to assess the presence of differentiated thyroid cancer.[3,4]

Ultrasound is considered a safe imaging modality and is commonly used to predict the malignancy of thyroid nodules.[1,3,4] Ultrasound is the first modality used to distinguish between benign and malignant nodules.[5,6] Over the last decade, research has indicated that some imaging features are more likely to be linked with thyroid cancer, including a solid composition, hypoechogenicity, an uneven or spiculated edge, a taller-than-wide form, and punctate echogenic foci.[2,3] Ultrasound characteristics such as microcalcifications, hypoechogenicity, and uneven edges have been linked to thyroid nodule malignancy; however, not all of them are highly predictive of the disease.[4]

Although a number of classification systems[7–9] have been developed to evaluate malignant thyroid nodules, ultrasound is a relatively subjective diagnostic method. Ultrasound parameters such as hypoechogenicity, calcification, composition, increased vascularity, and irregular margins are traditionally linked with the risk of malignancies, but they do not seem reliable enough to diagnose malignancy using ultrasound.[7–9]

Diagnostic ultrasound indices, such as sensitivity, specificity, and accuracy, have been shown to vary among observers and are therefore not the most reliable predictors of accurate and

This project was funded by the Deanship of Scientific Research (DSR) at King Abdul-Aziz University, Jeddah, under grant no. J: 48-142-1442. The authors, therefore, acknowledge with thanks DSR for technical and financial support.

The authors have no conflicts of interest to declare.

The datasets generated during and/or analyzed during the current study are not publicly available, but are available from the corresponding author on reasonable request. The data were collected at King Saud Medical City.

The study had been approved by the Institutional Review Board of King Saud Medical City (H1RI-22-Jun21-05)

[a] Department of Diagnostic Radiology, Faculty of Applied Medical Science, Imaging Unit, King Fahad Medical Research Centre, King Abdulaziz, Jeddah, Saudi Arabia, [b] Animal House Unit, King Fahad Medical Research Center, Faculty of Applied Medical Science, King Abdulaziz University, Jeddah, Saudi Arabia, [c] Smart Medical Imaging Research Group, King Abdulaziz University, Jeddah, Saudi Arabia, [d] Radiology Department, King Saud Medical City, Ministry of Health, Riyadh, Saudi Arabia.

**Correspondence: Jaber Alyami, Department of Diagnostic Radiology, Faculty of Applied Medical Science, King Abdul-Aziz University (KAU), Jeddah, Saudi Arabia (e-mail: jhalyami@kau.edu.sa).*

Copyright © 2022 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

How to cite this article: Alyami J, Almutairi FF, Aldoassary S, Albeshry A, Almontashri A, Abounassif M, Alamri M. Interobserver variability in ultrasound assessment of thyroid nodules. Medicine 2022;101:41(e31106).

Received: 17 January 2022 / Received in final form: 12 September 2022 / Accepted: 13 September 2022

http://dx.doi.org/10.1097/MD.0000000000031106

reproducible results. In ultrasound, observers may have different opinions when describing and interpreting thyroid lesions.[10] This can burden the already overloaded healthcare system and could lead to unnecessary use of FNAC examinations of thyroid nodules or surgical and treatment interventions.

It is understood that ultrasound is a very subjective method to use as an imaging modality for detecting thyroid nodules because its accuracy is based on the skill of the operator, and it is criticized for the variations caused by the observers.[11] Because of the high variability in interobserver ultrasound image interpretation, it is important to understand the reproducibility of the assessment of thyroid nodules to avoid unnecessary surgery or treatment intervention. Therefore, we conducted a cross-sectional retrospective analysis of recorded ultrasound images to evaluate the interobserver variability in ultrasound features of thyroid nodules.

2. Materials and Methods

2.1. Study design and setting

This retrospective study was performed between January 2021 and September 2021 in the ultrasound department of the King Saud Medical City (KSMC) in Riyadh, Saudi Arabia. Ultrasound imaging of the thyroid was performed using General Electric LOGIQ E9 scanner system. Patient data were retrieved from the hospital database and linked to the corresponding ultrasound images. The data were assigned randomly to 2 experienced radiologists for evaluation of the ultrasound images, and the different imaging features of the thyroid nodules were recorded. The final evaluation was performed and analyzed by the principal investigator using FNAC. The Institutional Review Board of the KSMC approved this study (Reference, H1RI-22-Jun 21-05).

2.2. Patients

The study included 195 ultrasound images of 75 thyroid nodules identified in 39 patients who underwent FNAC examinations at KSMC. Patients were required to have undergone both an ultrasound scan and FNAC examination and be aged 18 years or older to be included in the study. Patients under 18 years of age, and those who did not undergo either the ultrasound or the FNAC examination, were excluded.

2.3. Review of ultrasound images

Thyroid ultrasound images of 75 nodules were independently reviewed by 2 radiologists. Each radiologist had 6 years of experience in ultrasound imaging. The ultrasound images were evaluated blindly on the same liquid crystal display monitor, and the different ultrasound imaging features identified by the observers were recorded. They assessed nodules based on the 5 feature categories (echogenicity, shape; margin, vascularity, and calcification), in which ultrasound findings correspond to their association with malignancy (Fig. 1).

2.4. Statistical analysis

Statistical analyses were performed using SPSS version 26 (SPSS Corporation; Chicago, IL). Interobserver agreement was assessed using Cohen kappa statistic. The kappa scale was: values < 0.20 indicated slight agreement, 0.21–0.40 indicated fair agreement, 0.41–0.60 indicated moderate agreement, 0.61–0.80 indicated substantial agreement, and 0.81–1.00 indicated almost-perfect agreement.[12] For all statistics, the 95% confidence intervals (CIs) were also calculated. If a continuous variable was normally distributed, the mean and standard deviation are reported; the median and interquartile range (25th and 75th) are reported if the data were not normally distributed. A P value of < .05 was considered statistically significant.

3. Results

3.1. Study population

There were 33 (85%) females among the 39 patients in the study. The mean patient age was 45 ± 13 years (range: 24–73). Their mean body mass index was 26 ± 5 kg/m^2. There were 4 (10%) patients with malignancies (papillary carcinoma), 3 (8%) with diabetes, 1 (3%) with hypertension, and 2 (5%) with a family history of thyroid cancer.

3.2. Agreement for ultrasound features

Table 1 shows the measures of interobserver agreement between the 2 observers. The agreements in interpretation for echogenicity, shape, and margin were fair ($\kappa = 0.21–0.40$), whereas there were substantial agreements for vascularity and calcification ($\kappa = 0.62–0.78$). Apart from the margin-imaging feature, all other features (echogenicity, shape, vascularity, and calcification) showed significant differences ($P < .05$ each).

Table 2 shows the agreement between observers for ultrasound features reported in the current study compared with those stated in earlier studies of the same type. It was highest for vascularity and the presence/absence of calcification.

4. Discussion

Many classification guidelines for the assessment of thyroid nodules and consequent patient management have been developed. These include guidelines from the American College of Radiology (ACR), the Korean Thyroid Association/Korean Society of Thyroid Radiology, the European Thyroid Association, and the American Thyroid Association.[9,13–15] Classification guidelines help minimize cases of over-diagnosis or misdiagnosis by identifying whether thyroid nodules require either surgery or FNAC. However, the number of FNAC cases is increasing.[16,17]

Diagnosis is usually based on a composite set of ultrasound imaging features, including the margins, echogenicity, shape, vascularity, and presence or absence of calcifications. Therefore, it is important to define the risk associated with thyroid malignancy based on composite sets of ultrasound features. Since ultrasound analysis is dependent on the operator's ability to accurately identify ultrasound imaging features, a diagnosis can be affected by factors such as the observer's level of experience, the imaging system, and data acquisition. This study aimed to assess the interobserver variability in reporting ultrasound thyroid nodule imaging features and to compare the level of agreement in this study with other published studies.

Ultrasound is the imaging tool of choice for evaluating thyroid nodules, which are very common in clinical practice. The occurrence of thyroid nodules identified by ultrasound varies up to 68% of the population.[11] Several studies have investigated the capability of ultrasound to distinguish between benign thyroid nodules and malignant nodules based on ultrasound imaging features,[18,19] and many ultrasound imaging features have been reported to be associated with a high risk of thyroid malignancy. These include the presence of calcifications, hypoechogenicity, irregular borders, absence of halo signs, and central vascularity.[20] According to the ACR thyroid imaging reporting and data system (TI-RADS), thyroid nodule management should be performed based on ultrasound imaging features. Therefore, variability in ultrasound imaging features can lead to inconsistent management.

In this study, the interobserver variability of the ultrasound imaging features of shape, echogenicity, and margin showed

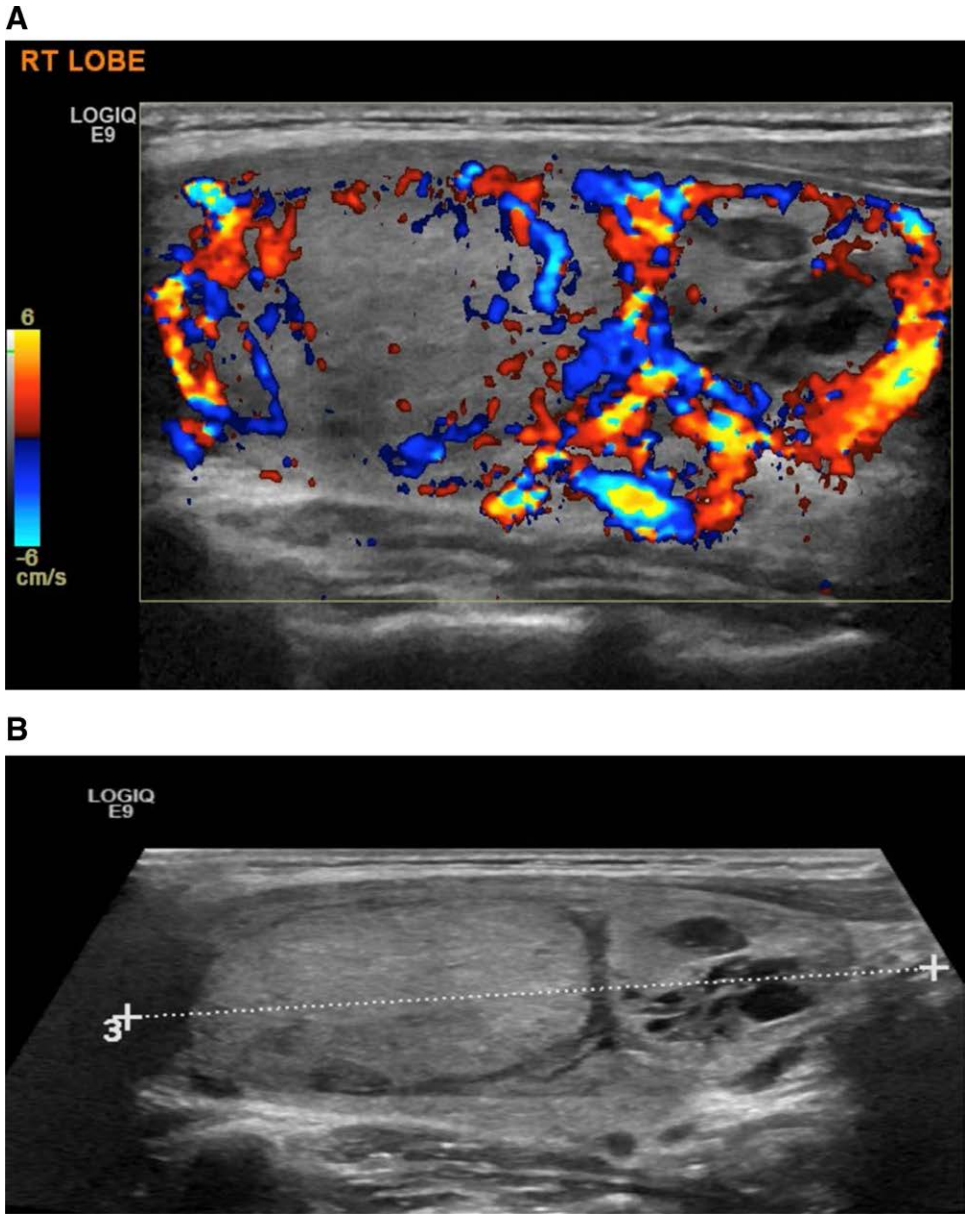

Figure 1. A 56-year-old woman with papillary thyroid carcinoma. Ultrasound images (A and B) show a heterogeneous nodule in the right thyroid lobe with high vascularity. This nodule was interpreted similarly by both observers and has a substantial agreement for US features.

Table 1

The measures of interobserver agreement between the two observers

			95% confidence interval of the difference		
	Interobserver agreement (%)	Kappa	Lower	Upper	*P* value
Echogenicity	0.345	0.36	–0.454	0.044	.032
Shape	0.452	0.40	–0.326	–0.033	.004
Margin	0.24	0.21	–0.34	–0.020	.136
Vascularity	0.83	0.62	–0.084	0.34	<.001
Calcification	0.917	0.78	–0.116	0.065	<.001

Kappa = Cohen kappa coefficient, *P* value = the probability value.

a fair level of agreement (κ = 0.21–0.40), whereas vascularity and calcification had more substantial agreement with κ = 0.62–0.78. Our results are consistent with those reported by Moon et al[7] for shape, echogenicity, and calcifications. Moreover, a study reported fair interobserver agreement for margin ultrasound features with a simple percentage of 28% and a kappa value of 0.25, similar to our study.[21] In contrast, relatively low kappa values were reported for margin features that had a poor agreement between observers (κ = 0.14) compared to our study; while, in the same study, echogenicity had a

Table 2

The agreement between observers for ultrasound features

	Current study	Kim, 2012[13]	Koltin, 2016[14]	Lim-Dunham, 2017[15]	Wienke, 2003[16]
Nodules	75	80	27	39	70
Statistics	Kappa	Kappa	Kappa	Kappa	Kappa
Observers	2 experienced radiologists	7 resident radiologists, 2 different units	3 experienced radiologists	2 experienced radiologists	2 radiologists
Echogenicity	0.36	0.5	0.46	0.54	0.37
Shape	0.40	0.57	N/A	0.29	N/A
Margin	0.21	0.49	0.58	0.6	0.13
Vascularity	0.62	N/A	0.18	0.76	0.75
Calcification	0.78	0.62	N/A	N/A	0.91

Kappa = Cohen kappa coefficient.

similar level of agreement ($\kappa = 0.33$ vs 0.36).[21] Similarly, Kim et al (2010) showed moderate agreement for echogenicity, calcification, and fair agreement for margin features;[22] however, the agreement was assessed based on 5 observers compared to 2 in our study. Grani et al showed interobserver agreement consistent with our study, which showed moderate agreement for both echogenicity and margin and substantial agreement for calcification.[23] However, their study was limited to only benign cases.

The level of agreement in this study is consistent with what has been reported in other studies for echogenicity, vascularity, and calcification, but differs for margin.[24] This difference might be attributable to the level of observer experience or the number of nodules. In studies that had a high level of agreement for the margin, the assessments involved 3 or 7 observers.[25,26] In addition, the number of nodules in the study by Koltin et al[28] was relatively small compared to that in the current study.

The current study has several limitations. First, the study design was retrospective. All patients involved had undergone ultrasound-guided FNAC; therefore, bias may have existed in the selection of patients included in the study. Second, the observers assessed static images retrieved from the hospital database and did not perform or evaluate the ultrasound scans themselves. Thus, variability could arise because the observers could not benefit from specific ultrasound operational features. Third, intraobserver variability was not evaluated. Finally, the sample size used in the current study was relatively small. This could affect the variability and, therefore, the level of agreement.

A systematic approach to thyroid ultrasound training and familiarity with ACR TI-RADS guidelines are recommended for improving interobserver agreement. Such systematic training may lead to more-appropriate decisions, with better consistency among observers, regarding whether to perform FNAC to follow-up and evaluate the thyroid nodules. Therefore, a multicenter study with different practices is recommended to assess the variability in ultrasound characteristics among observers.

In conclusion, this study found wide variability in the evaluation of individual ultrasound features. Therefore, following a systematic approach for evaluating thyroid nodules may improve the interobserver agreement and minimize the level of variability. This study emphasizes the importance of dedicated standard training modules for US Scan observers.

Acknowledgements

This project was funded by the Deanship of Scientific Research (DSR) at King Abdul-Aziz University, Jeddah, under grant no. J: 48-142-1442. The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Author contributions

Conceptualization: Fahad F. Almutairi, Jaber Alyami, Mazen Abounassif.
Data curation: Ali Almontashri, Amani Albeshry, Fahad F. Almutairi, Majed Alamri, Mazen Abounassif.
Formal analysis: Ali Almontashri, Fahad F. Almutairi, Majed Alamri, Mazen Abounassif, Sultan Aldoassary.
Funding acquisition: Jaber Alyami.
Investigation: Ali Almontashri, Amani Albeshry, Majed Alamri, Mazen Abounassif, Sultan Aldoassary.
Methodology: Ali Almontashri, Fahad F. Almutairi, Majed Alamri, Mazen Abounassif, Sultan Aldoassary.
Project administration: Jaber Alyami, Majed Alamri.
Resources: Amani Albeshry, Sultan Aldoassary.
Software: Ali Almontashri, Fahad F. Almutairi, Majed Alamri.
Supervision: Ali Almontashri, Jaber Alyami, Mazen Abounassif.
Validation: Ali Almontashri, Fahad F. Almutairi, Mazen Abounassif, Sultan Aldoassary.
Visualization: Ali Almontashri, Jaber Alyami.
Writing – original draft: Fahad F. Almutairi, Jaber Alyami.
Writing – review & editing: Jaber Alyami.

References

[1] Hegedus L, Bonnema SJ, Bennedbaek FN. Management of simple nodular goiter: current status and future perspectives. Endocr Rev. 2003;24:102–32.
[2] Hoang JK, Lee WK, Lee M, et al. US Features of thyroid malignancy: pearls and pitfalls. Radiographics. 2007;27:847–60; discussion 861.
[3] Frates MC, Benson CB, Charboneau JW, et al. Management of thyroid nodules detected at US: society of radiologists in ultrasound consensus conference statement. Radiology. 2005;237:794–800.
[4] Hoang JK, Middleton WD, Farjat AE, et al. Interobserver variability of sonographic features used in the American College of Radiology Thyroid Imaging Reporting and Data System. Am J Roentgenol. 2018;211:162–7.
[5] Xie C, Cox P, Taylor N, et al. Ultrasonography of thyroid nodules: a pictorial review. Insights Imag. 2016;7:77–86.
[6] Jeong EJ, Chung SR, Baek JH, et al. A comparison of ultrasound-guided fine needle aspiration versus core needle biopsy for thyroid nodules: pain, tolerability, and complications. Endocrinol Metab. 2018;33:114–20.
[7] Moon W-J, Jung SL, Lee JH, et al. Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. Radiology. 2008;247:762–70.
[8] Chung R, Rosenkrantz AB, Bennett GL, et al. Interreader concordance of the TI-RADS: impact of radiologist experience. Am J Roentgenol. 2020;214:1152–7.
[9] Tessler FN, Middleton WD, Grant EG, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. J Am Coll Radiol. 2017;14:587–95.
[10] Park CS, Kim SH, Jung SL, et al. Observer variability in the sonographic evaluation of thyroid nodules. J Clin Ultrasound. 2010;38:287–93.
[11] Choi SH, Kim E-K, Kwak JY, et al. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid. 2010;20:167–72.
[12] Kursuncu U, Gaur M, Castillo C, et al. Modeling islamist extremist communications on social media using contextual dimensions: religion, ideology, and hate. Proc ACM Human Comp Inter. 2019;3:1–22.
[13] Haugen BR, Alexander EK, Bible KC, 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.
[14] Shin JH, Baek JH, Chung J, 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.
[15] Russ G, Bonnema SJ, Erdogan MF, et al. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. Eur Thyroid J. 2017;6:225–37.
[16] Ng DL, van Zante A, Griffin A, et al. A large thyroid fine needle aspiration biopsy cohort with long-term population-based follow-up. Thyroid. 2021;31:1086–95.
[17] Zhu Y, Song Y, Xu G, et al. Causes of misdiagnoses by thyroid fine-needle aspiration cytology (FNAC): our experience and a systematic review. Diagn Pathol. 2020;15:1–8.
[18] Rowe ME, Osorio M, Likhterov I, et al. Evaluation of ultrasound reporting for thyroid cancer diagnosis and surveillance. Head Neck. 2017;39:1756–60.
[19] Wong K, Ahuja AT. Ultrasound of thyroid cancer. Cancer Imag. 2005;5:157.
[20] Remonti LR, Kramer CK, Leitao CB, et al. Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. Thyroid. 2015;25:538–50.
[21] Guth S, Theune U, Aberle J, et al. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest. 2009;39:699–706.
[22] Kim SH, Park CS, Jung SL, et al. Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules. Korean J Radiol. 2010;11:149–55.
[23] Grani G, Lamartina L, Cantisani V, et al. Interobserver agreement of various thyroid imaging reporting and data systems. Endocr Connect. 2018;7:1–7.
[24] Wienke JR, Chong WK, Fielding JR, et al. Sonographic features of benign thyroid nodules: interobserver reliability and overlap with malignancy. J Ultrasound Med. 2003;22:1027–31.
[25] Koltin D, O'Gorman CS, Murphy A, et al. Pediatric thyroid nodules: ultrasonographic characteristics and inter-observer variability in prediction of malignancy. J Pediatr Endocrinol Metab. 2016;29:789–94.
[26] Lim-Dunham JE, Erdem Toslak I, Alsabban K, et al. Ultrasound risk stratification for malignancy using the 2015 American Thyroid Association Management Guidelines for Children with Thyroid Nodules and Differentiated Thyroid Cancer. Pediatr Radiol. 2017;47:429–36.
[27] Koltin D, et al. Pediatric thyroid nodules: ultrasonographic characteristics and inter-observer variability in prediction of malignancy. J Pediatr Endocrinol Metab. 2016;29:789–94.