The Oncologist®

Endocrinology

A Multidisciplinary Head-to-Head Comparison of American College of Radiology Thyroid Imaging and Reporting Data System and American Thyroid Association Ultrasound Risk Stratification Systems

Bernice L. Huang,[a] Susana A. Ebner,[b] Jasnit S. Makkar,[c] Stuart Bentley-Hibbert,[c] Robert J. McConnell,[b,d] James A. Lee,[d,e]
Elizabeth M. Hecht,[c] Jennifer H. Kuo[d,e]

[a]Department of Surgery, [b]Division of Endocrinology, Department of Medicine, [c]Department of Radiology, and [d]The Columbia Thyroid Center, Columbia University Medical Center, New York, New York, USA; [e]Division of GI/Endocrine Surgery, Department of Surgery, Columbia University Irving Medical Center, New York, New York, USA

Disclosures of potential conflicts of interest may be found at the end of this article.

Key Words. Thyroid nodule • Thyroid cancer • Ultrasonography • Interobserver variability

Abstract

Background. Ultrasound plays a critical role in evaluating thyroid nodules. We compared the performance of the two most popular ultrasound malignancy risk stratification systems, the 2015 American Thyroid Association (ATA) guidelines and the American College of Radiology Thyroid Imaging and Reporting Data System (ACR TI-RADS).

Materials and Methods. We retrospectively identified 250 thyroid nodules that were surgically removed from 137 patients. Their ultrasound images were independently rated using both ATA and ACR TI-RADS by six raters with expertise in ultrasound interpretation. For each system, we generated a receiver operating characteristic curve and calculated the area under the curve (AUC).

Results. Sixty-five (26%) nodules were malignant. There was "fair agreement" among raters for both ATA and ACR TI-RADS. Our observed malignancy risks for ATA and ACR TI-RADS categories were similar to expected risk thresholds with a few notable exceptions including the intermediate ATA risk category and the three highest risk categories for ACR TI-RADS. Biopsy of 226 of the 250 nodules would be indicated by ATA guidelines based on nodule size and mean ATA rating. One hundred forty-six nodules would be biopsied based on ACR TI-RADS. The sensitivity, specificity, and negative and positive predictive values were 92%, 10%, 79%, and 27%, respectively, for ATA and 74%, 47%, 84%, and 33%, respectively, for ACR TI-RADS. The AUC for ATA was 0.734 and for ACR TI-RADS was 0.718.

Conclusion. Although both systems demonstrated good diagnostic performance, ATA guidelines resulted in a greater number of thyroid biopsies and exhibited more consistent malignancy risk prediction for higher risk categories. ***The Oncologist*** 2020;25:398–403

Implications for Practice: With the rising incidence of thyroid nodules, the need for accurate detection of malignancy is important to avoid the overtreatment of benign nodules. Ultrasonography is one of the key tools for the evaluation of thyroid nodules, although the use of many different ultrasound risk stratification systems is a hindrance to clinical collaboration in everyday practice and the comparison of data in research. The first step toward the development of a universal thyroid nodule ultrasound malignancy risk stratification system is to better understand the strengths and weaknesses of the current systems in use.

Introduction

Thyroid nodules are a common finding, with a prevalence of 30%–67% in the general population according to ultrasound screening and autopsy studies [1, 2]. The incidence of thyroid nodules and thyroid cancers has grown substantially over the past 2 decades, in part due to the detection of small asymptomatic thyroid nodules as a result of

Correspondence: Jennifer H. Kuo, M.D., M.S., Division of GI/Endocrine Surgery, Dept. of Surgery, Columbia University Irving Medical Center, 161 Fort Washington Ave., Suite 829, New York, New York 10032, USA. Telephone: 212-305-6969; e-mail: jhk2029@cumc.columbia.edu Received May 11, 2019; accepted for publication October 17, 2019; published Online First on November 19, 2019. http://dx.doi.org/10.1634/theoncologist.2019-0362

advancements in medical surveillance and increased use of imaging (ultrasonography, computed tomography, magnetic resonance imaging) [3]. However, the majority of nodules are benign, with only an estimated 7%–15% exhibiting malignancy [4]. In order to avoid overtreatment of thyroid nodules, it is important to be able to accurately distinguish benign from malignant nodules prior to proceeding with surgical excision. One of the most important diagnostic tools we have toward that end is ultrasonography.

Thyroid ultrasonography is cost-effective and, in conjunction with demographic and clinical factors, helps determine which nodules warrant further investigation with fine needle aspiration biopsy or additional surveillance. Although no single ultrasound feature has been identified to accurately diagnose malignancy in thyroid nodules [5], a combination of suspicious features has proved to be helpful [6, 7]. Over the past few years, multiple different risk stratification systems based on the combination of ultrasound characteristics of thyroid nodules have been developed worldwide [6, 8–11]. In general, the different systems focus on the same ultrasound characteristics that have been associated with malignancy including echogenicity, shape, margins, and presence of calcifications. However, they each have a different approach to assigning malignancy risk according to these characteristics. In the U.S., the two most commonly used systems are the 2015 American Thyroid Association (ATA) management guidelines [4] and the 2017 American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) [12]. The ATA guidelines are a qualitative system that stratifies nodules into five different risk categories, each defined by a constellation of sonographic findings. On the other hand, the ACR TI-RADS also has five different risk categories but assigns points to specific sonographic findings within broader categories of ultrasound characteristics and determines risk stratification quantitatively based on the summation of those points. Both systems have been independently validated for use in predicting malignancy [13, 14], although few studies have directly compared the performance of the two systems. Our goal for this study was to perform a head-to-head, multidisciplinary comparison of the ATA and ACR TI-RADS systems based on interobserver agreement, accuracy of malignancy risk prediction, and overall diagnostic performance in predicting malignancy.

Materials and Methods

Patients

After obtaining institutional review board approval (AAAD4780), we retrospectively identified patients with thyroid nodules who underwent thyroid lobectomy or total thyroidectomy at Columbia University Irving Medical Center from 2017 to 2018. Only patients with full ultrasound images of their dominant nodules with clear correlating surgical pathology findings were included. A total of 250 thyroid nodules from 137 patients were analyzed.

Ultrasound Image Interpretation

Ultrasound images were obtained in the radiology department by ultrasound technicians or in clinic by an endocrine surgeon. A 5- to 12-MHz linear probe was used on three different machines (GE Logiq E9 [GE Healthcare, Milwaukee, WI], ProSound Alpha 6 [Hitachi Aloka Medical, Tokyo, Japan], Terason uSmart 3200T [Terason Ultrasound, Burlington, MA]), all Food and Drug Administration approved for use in neck ultrasonography. Representative preoperative sonographic JPEG images (transverse and longitudinal views) for all nodules were collated into a blinded online survey by an investigator not involved in image interpretation. Images were independently reviewed by a group of multidisciplinary raters including three board-certified fellowship-trained radiologists (with 3 [J.M.], 11 [S.B.H.], and 16 [E.H.] years of experience), two endocrinologists (with 17 [S.E.] and 20 [R.M.] years of experience), and one endocrine surgeon (with 6 years of experience [J.K.]), all experts in interpreting neck ultrasonography. Both endocrinologists and the endocrine surgeon in this study individually perform more than 1,000 diagnostic ultrasound evaluations yearly, and S.E. and J.K. have obtained Endocrine Certification in Neck Ultrasound by the American Association of Clinical Endocrinologists. Raters individually assigned malignancy risk categories for the nodules in accordance with the ATA and/or the ACR TI-RADS systems. According to rater preference, five of the raters evaluated nodules using both rating systems and one rater (R.M.) evaluated nodules using the ATA system only.

Table 1. Patient demographics and thyroid nodule characteristics

Demographics and nodule characteristics	*n* (%)
Age, median (IQR), years	58 (46–69)
Sex	
Female	212 (85)
Male	48 (15)
Ethnicity	
White	117 (47)
Hispanic	47 (19)
Black	39 (16)
Asian	8 (3)
Unknown	39 (16)
Nodule size, median (IQR), cm	2.2 (1.6–3.3)
NIFTP	6 (2)
Malignant	65 (26)
Papillary	56 (86)
Follicular	7 (11)
Hurthle	3 (43)
Medullary	2 (3)

Abbreviations: IQR, interquartile range; NIFTP, noninvasive follicular thyroid neoplasm with papillary-like nuclear features.

Statistical Analysis

Descriptive statistics of demographic information and surgical pathology results were performed. Continuous variables were expressed as median and interquartile range. Interobserver agreement was assessed using Fleiss' kappa coefficient [15] and stratified by rater specialty (clinical vs. radiology) and average ATA or ACR TI-RADS rating. Observed and expected malignancy risks for each risk category in the

Table 2. Interobserver agreement for each risk stratification system by observer specialty and average ATA or ACR TI-RADS rating

Risk stratification system	Fleiss' κ[a] by specialty			Fleiss' κ by average risk strata				
	All raters	Clinicians	Radiologists	1	2	3	4	5
ATA	0.281	0.313	0.304	—[b]	0.104	0.129	0.075	0.069
ACR TI-RADS	0.313	0.274	0.360	−0.067	0.090	0.126	0.065	−0.112

[a] Interpretation of κ values: <0 = poor agreement, 0.0–0.20 = slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, 0.81–1.0 = almost perfect agreement [27].
[b] Limited sample size. Unable to calculate κ value.
Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging and Reporting Data System; ATA, 2015 American Thyroid Association Guidelines.

Table 3. Comparison of observed and expected malignancy risks by risk stratum

Risk stratification	All nodules	Malignant nodules	Observed malignancy, %	Expected malignancy,[a] %	*p* value
ATA risk stratification					
Benign	1	0	0	<1	.99
Very low suspicion	14	0	0	<3	.99
Low suspicion	127	18	14	5–10	.14
Intermediate suspicion	89	32	36	10–20	<.001
High suspicion	19	15	79	>70–90	.32
ACR TI-RADS risk stratification					
TR1 - benign	4	0	0	0	.99
TR2 - not suspicious	30	1	3	<2	.99
TR3 - mildly suspicious	103	17	17	≤5	<.001
TR4 - moderately suspicious	88	34	39	5.1–20	<.001
TR5 - highly suspicious	25	13	52	>20	<.001

[a] Expected malignancy rates as previously reported [14, 16].
Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging and Reporting Data System; ATA, 2015 American Thyroid Association Guidelines

Table 4. Diagnostic performance based on theoretical biopsy criteria

Risk stratification system	Nodules biopsied, *n*	Sensitivity	Specificity	NPV	PPV
ATA	226	92%	10%	79%	27%
ACR TI-RADS	146	74%	47%	84%	33%

Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging and Reporting Data System; ATA, 2015 American Thyroid Association Guidelines; NPV, negative predictive value; PPV, positive predictive value.

stratification systems were compared using chi-square test with Bonferroni correction for multiple comparisons. The upper limit of expected malignancy risks according to guidelines was used for comparison [4, 16]. Assessment of diagnostic performance was performed in a stepwise fashion. First, we determined which nodules would meet biopsy criteria according to each system based on their average risk category and nodule size [4, 12]. This was compared with surgical pathology results to determine the sensitivity, specificity, positive predictive value, and negative predictive value (NPV) of each system. Receiver operating characteristic (ROC) curves were generated and the area under the curve (AUC) was calculated for each system using linear predictors of the curve. The AUCs for the two systems were compared using a bootstrap method for correlated ROC curves [17]. Statistical analysis was performed using R software 3.4.3 (R Foundation for Statistical Computing, Vienna, Austria) [18–20].

Results

Patient Demographics

Table 1 summarizes patient demographics and nodule characteristics. The median age of our cohort was 58 years (46–69). The majority were women (*n* = 212 [85%]). The median nodule size was 2.2 cm (1.6–3.3). Malignant nodules accounted for 65 (26%) of all nodules. The majority of malignant nodules were papillary thyroid carcinomas (86%). A small number of surgical specimens were found to contain incidental papillary microcarcinomas (<1 cm, *n* = 9 [4%]) that

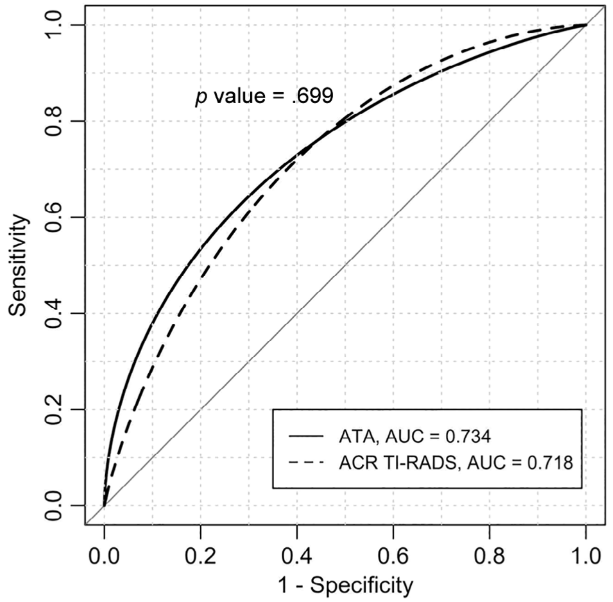

Figure 1. Receiver operating characteristic curves of ATA and ACR TI-RADS risk stratification systems.
Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging and Reporting Data System; ATA, 2015 American Thyroid Association Guidelines; AUC, area under the curve.

were not the dominant nodules being evaluated. Six nodules (2%) were noninvasive follicular thyroid neoplasm with papillary-like nuclear features and were also classified as benign nodules.

Interobserver Agreement

Table 2 summarizes the Fleiss' κ values for interobserver agreement. When accounting for all raters, κ for ATA was 0.281 and for ACR TI-RADS was 0.313 (fair agreement for both). When stratified by rater specialty, κ for ATA was 0.313 among endocrinologists/endocrine surgeon and was 0.304 among radiologists. κ for ACR TI-RADS was 0.274 among endocrinologists/endocrine surgeon and was 0.360 among radiologists. When analyzed by average ATA rating, κ for ATA ratings from very low to high suspicion were 0.104, 0.12, 0.075, and 0.069. Interobserver agreement could not be measured for the ATA benign category because of limited sample size (n = 1). When analyzed by average ACR TI-RADS rating, κ for TIRADS ratings from benign to highly suspicious were − 0.067, 0.090, 0.126, 0.065, and − 0.112.

Accuracy of Risk Strata and Expected Malignancy Risks

The total number of nodules that fell into each stratum of the ATA and ACR TI-RADS systems according to the mean total rater score, calculated observed malignancy risk, and expected malignancy risk of each stratum are shown in Table 3. There was no significant difference in observed malignancy risk versus expected for four of the five strata of the ATA classification system. However, our calculated observed malignancy risk of 36% for the intermediate suspicion category was significantly higher than the expected 10%–20% (p < .001). Our calculated observed malignancy risks were significantly higher than the expected malignancy risks for three of the five strata in the ACR TI-RADS system: 17% versus ≤5% (p < .001) for TR3, 39% versus 5%–20% (p < .001) for TR4, and 52% versus >20% for TR5.

Diagnostic Performance

Table 4 summarizes the diagnostic performance of both risk stratification systems based on the theoretical application of biopsy criteria for each system and the surgical pathology findings of benign versus malignant nodules. The sensitivity, specificity, positive predictive value, and negative predictive value for ATA were 92%, 10%, 79%, and 27%, respectively, and for ACR TI-RADS were 74%, 47%, 84%, and 33%, respectively.

Figure 1 illustrates the ROC curve for both ATA and ACR TI-RADS risk stratification systems. The AUC for ATA was 0.734 (confidence interval [CI]: 0.663–0.804). The AUC for ACR TI-RADS was 0.718 (CI: 0.643–0.784). There was no statistically significant difference between the AUCs for both systems (p = .699).

Discussion

Although there have been a few studies that have compared different risk stratification systems [21–24], and in particular, the ATA and ACR-TIRADS systems [25, 26], to our knowledge, this is the first study to evaluate the performance of these two systems using raters from multiple clinical disciplines. We found that both ATA and ACR TI-RADS risk stratifications systems exhibited similar diagnostic performance and interobserver agreement, although ATA guidelines were more accurate in predicting the malignancy risk for each risk stratum in our study population.

Both ATA and ACR TI-RADS risk stratification systems demonstrated "fair" interobserver agreement overall [27]. Not surprisingly, when stratified by discipline, radiologists had greater interobserver agreement using ACR TI-RADS than when using the ATA system. Similarly, endocrinologists/endocrine surgeon had better interobserver agreement using the ATA system than the ACR TI-RADS system. Although we can only speculate on the etiology of this difference, one possible reason could be that radiologists are trained to use the ACR TI-RADS system and endocrinologists/endocrine surgeon are trained to use the ATA system. Our findings are in line with prior studies using Fleiss' kappa to evaluate interobserver agreement. In 2018, Hoang et al. found a κ value of 0.35 for ACR TI-RADS grading in the evaluation of 100 thyroid nodules by eight board-certified radiologists of different levels of training and practice settings [28]. Grani et al. compared interobserver concordance between two endocrinologists who interpreted ultrasound images from 501 nodules based on a set of characteristics that were later applied to obtain ratings according to different risk stratification systems [29]. For both ATA and ACR TI-RADS systems, they obtained a Krippendorff alpha value of 0.49. Interestingly, after having a joint session to discuss ratings and reach consensus readings for this first set of nodules, the two raters were asked to independently review a separate set of 554 nodules, which resulted in an improvement in Krippendorff alpha to 0.57

(ACR TI-RADS) and 0.65 (ATA). Their findings indicate that both systems exhibit similar degrees of interobserver agreement and that standardized training in each system may further improve interobserver agreement.

When we compared the expected rates of malignancy for each stratum in the risk stratification systems, we found some discrepancies. In particular, we noted that we had a much higher risk of malignancy than expected for the intermediate risk stratum in the ATA guidelines (36% vs. 10%–20%) as well as the ACR TI-RADS 3–5 categories. A similar study was performed by Gao et al. to compare the TI-RADS system proposed by Kwak (Kwak-TIRADS) with ATA and ACR TI-RADS systems in 2,544 nodules that were also surgically excised and had an even higher overall malignancy rate of 66.1% compared with ours. They, too, found significantly higher observed rates of malignancy for the intermediate ATA risk stratum and the higher ACR TI-RADS risk strata in comparison with expected malignancy thresholds [25]. Two possible explanations for this discrepancy are as follows. First, it is possible that our cohort, given that they had been referred to a surgical clinic, had a higher incidence of other confounding risk factors for malignancy, such as family history of thyroid cancer, which could have impacted the malignancy rates, although one would expect this to have an equal effect across the board for all risk strata. A second possible explanation is the high proportion of follicular cancers that were included in our study (11% of malignant nodules) in comparison with 1% of malignant nodules in the ACR TI-RADS validation study [14] and 5% of malignant nodules in the ATA validation study [13]. This is supported by the fact that our intermediate ATA risk stratum and ACR TI-RADS 3–4 risk strata had higher proportions of follicular cancers in their malignant nodules (12%–27%) in comparison with the other risk strata (0%–9%), with the exception of ACR TI-RADS 2, in which the only malignant nodule was a follicular cancer, and ACR TI-RADS 5, in which none of the 13 malignant nodules were follicular cancer. Regardless of their cause, these discrepancies are important to note because the reported risk of malignancy of a thyroid nodule based on its risk stratification can significantly influence the treatment preferences of a patient and provider, which are often just as important as treatment recommendations in clinical guidelines in the final decision regarding the individual treatment plan for a thyroid nodule. It should be stressed that the ultrasound malignancy risk estimations are only a single data point that must be taken in the context of all other demographic and clinical information when applied toward treatment decisions.

In comparing the diagnostic performance of ATA and ACR TI-RADS, we found that ATA had a higher sensitivity and NPV than ACR TI-RADS, which had higher specificity than ATA. This resulted in a higher number of theoretical biopsies indicated based on application of ATA criteria in comparison with ACR TI-RADS. However, the AUCs for the ROC curves for both systems were not statistically different, indicating similar overall performance in predicting malignancy. These findings are in line with other studies comparing both ATA and ACR TI-RADS. In the study by Gao et al., they also found that ACR TI-RADS had higher specificity (79.7%, *p* < .05) and ATA had higher sensitivity (95.5%, *p* < .05), although their ultimate conclusion was that KWAK-TIRADS had better diagnostic performance in differentiating nodules >1 cm (AUC: 0.92, *p* < .01) in comparison with the other two systems. Another study by Ha et al. compared seven different society guidelines, including ATA and ACR TI-RADS, in the evaluation of 2,000 thyroid nodules. Their final diagnosis of malignancy was determined with a combination of both surgical pathology and fine needle aspiration or core needle biopsy in cases in which patients did not undergo surgery. Again, they noted that ATA had higher sensitivity (89.6% vs. 74.7%) and ACR TI-RADS had higher specificity (67.3% vs. 33.2%). Out of all seven systems, ACR TI-RADS resulted in the lowest rate of unnecessary biopsies (25.3%), compared with a rate of 51.7% for ATA. In a subsequent study, Ha et al. compared just ATA, ACR TI-RADS, and the 2016 Korean Thyroid Association guidelines in the evaluation of 902 additional nodules and found, similarly, that ATA had higher sensitivity than ACR TI-RADS (95.0% vs. 80.2%, *p* = .001) but also had lower specificity (38.1% vs. 68.9%, *p* < .001), and ACR TI-RADS had the lowest rate of unnecessary biopsies (25.8%) [30]. The clinical implication of this difference is hard to define. Although the use of the ATA risk stratification system results in a higher detection of thyroid cancer, this also results in treatment of those cancers, and potential *overtreatment* of those cancers. In our current era of less aggressive treatment for this mostly indolent disease, the utility of increased detection remains an ambiguous benefit.

There are several limitations to our study. This is a retrospective analysis of a surgical cohort that traditionally has had a higher rate of malignancy than the general thyroid nodule population in an institution. However, our malignancy rate of 26% in this study cohort is on par with the general rate of malignancy of thyroid nodules seen at our institution (24%). This study also only included preselected dimensional sonographic images. Additional images through the entire nodule, review of the background parenchyma, and real-time video clips of the sonographic examination may have changed interpretation of some selected features and categorization. We also have a relatively small cohort of 250 nodules included in the study, with a small number of nodules in specific strata limiting substratification analyses. Despite these limitations, we believe that our data add insight into the utility of these tests and the practice preferences of the clinicians who use them.

Conclusion

In the handful of studies to date that have compared different thyroid nodule ultrasound risk stratification systems [21–26, 30, 31], it remains unclear that any single system outperforms the other with respect to statistical or clinical significance. However, these studies have shown that most of these systems have demonstrated relatively accurate predictions of malignancy, with high sensitivity and NPV. Unfortunately, having so many different rating systems in use can be confusing for both patients and providers. It can also hinder the aggregation and comparison of data in research on thyroid nodules. Moving forward, continuing to work across disciplines and internationally to build consensus guidelines for risk stratification of thyroid nodules is critical because it would provide a common lexicon and framework for multicenter, prospective trials that would benefit patient care and facilitate training of physicians who clinically interpret thyroid ultrasound.

Author Contributions

Conception/design: Bernice L. Huang, James A. Lee, Elizabeth M. Hecht, Jennifer H. Kuo
Provision of study material or patients: Susana A. Ebner, Jasnit S. Makkar, Stuart Bentley-Hibbert, Robert J. McConnell, James A. Lee, Elizabeth M. Hecht, Jennifer H. Kuo
Collection and/or assembly of data: Bernice L. Huang, Jennifer H. Kuo
Data analysis and interpretation: Bernice L. Huang, Jennifer H. Kuo
Manuscript writing: Bernice L. Huang, Susana A. Ebner, Jasnit S. Makkar, Stuart Bentley-Hibbert, Robert J. McConnell, James A. Lee, Elizabeth M. Hecht, Jennifer H. Kuo
Final approval of manuscript: Bernice L. Huang, Susana A. Ebner, Jasnit S. Makkar, Stuart Bentley-Hibbert, Robert J. McConnell, James A. Lee, Elizabeth M. Hecht, Jennifer H. Kuo

Disclosures

The authors indicated no financial relationships.

References

1. 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.

2. Mortensen JD, Woolner LB, Bennett WA. Gross and microscopic findings in clinically normal thyroid glands. J Clin Endocrinol Metab 1955;15: 1270–1280.

3. Vaccarella S, Dal Maso L, Laversanne M et al. The impact of diagnostic changes on the rise in thyroid cancer incidence: A population-based study in selected high-resource countries. Thyroid 2015;25: 1127–1136.

4. 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.

5. Brito JP, Gionfriddo MR, Al Nofal A et al. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: Systematic review and meta-analysis. J Clin Endocrinol Metab 2014;99:1253–1263.

6. Kwak JY, Yoon JH, Moon HJ et al. Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. Radiology 2011;260: 892–899.

7. Smith-Bindman R, Lebda P, Feldstein VA et al. Risk of thyroid cancer based on thyroid ultrasound imaging characteristics: Results of a population-based study. JAMA Intern Med 2013;173: 1788–1796.

8. Horvath E, Rossi R, Franco C et al. An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. J Clin Endocrinol Metab 2009;94:1748–1751.

9. 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;6:225–237.

10. Gharib H, Papini E, Garber JR et al. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules - 2016 update. Endocr Pract 2016;22: 622–639.

11. Perros P, Boelaert K, Colley S et al. Guidelines for the management of thyroid cancer. Clin Endocrinol (Oxf) 2014;81(suppl 1):1–122.

12. 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;14:587–595.

13. Tang AL, Falciglia M, Yang H et al. Validation of American Thyroid Association ultrasound risk assessment of thyroid nodules selected for ultrasound fine-needle aspiration. Thyroid 2017;27: 1077–1082.

14. Zheng Y, Xu S, Kang H et al. A single-center retrospective validation study of the American College of Radiology Thyroid Imaging Reporting and Data System. Ultrasound Q 2018;34:77–83.

15. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76: 378–382.

16. Middleton WD, Teefey SA, Reading CC et al. Multiinstitutional analysis of thyroid nodule risk stratification using the American College of Radiology Thyroid Imaging Reporting and Data System. AJR Am J Roentgenol 2017;208:1331–1341.

17. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–843.

18. R: A language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2017.

19. irr: various coefficients of interrater reliability and agreement. R package [computer program]. Version 0.842012.

20. Robin X, Turck N, Hainard A et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12:77.

21. Yoon JH, Lee HS, Kim EK et al. Malignancy risk stratification of thyroid nodules: Comparison between the Thyroid Imaging Reporting and Data System and the 2014 American Thyroid Association Management Guidelines. Radiology 2016;278:917–924.

22. Chng CL, Tan HC, Too CW et al. Diagnostic performance of ATA, BTA and TIRADS sonographic patterns in the prediction of malignancy in histologically proven thyroid nodules. Singapore Med J 2018;59:578–583.

23. Ha EJ, Na DG, Baek JH et al. US fine-needle aspiration biopsy for thyroid malignancy: Diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. Radiology 2018; 287:893–900.

24. Macedo BM, Izquierdo RF, Golbert L et al. Reliability of Thyroid Imaging Reporting and Data System (TI-RADS), and ultrasonographic classification of the American Thyroid Association (ATA) in differentiating benign from malignant thyroid nodules. Arch Endocrinol Metab 2018;62:131–138.

25. Gao L, Xi X, Jiang Y et al. Comparison among TIRADS (ACR TI-RADS and KWAK- TI-RADS) and 2015 ATA Guidelines in the diagnostic efficiency of thyroid nodules. Endocrine 2019;64:90–96.

26. Middleton WD, Teefey SA, Reading CC et al. Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association Guidelines. AJR Am J Roentgenol 2018;210:1148–1154.

27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.

28. 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. AJR Am J Roentgenol 2018;211:162–167.

29. Grani G, Lamartina L, Cantisani V et al. Interobserver agreement of various thyroid imaging reporting and data systems. Endocr Connect 2018; 7:1–7.

30. Ha EJ, Na DG, Moon WJ et al. Diagnostic performance of ultrasound-based risk-stratification systems for thyroid nodules: Comparison of the 2015 American Thyroid Association Guidelines with the 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology and 2017 American Congress of Radiology Guidelines. Thyroid 2018; 28:1532–1537.

31. Maino F, Forleo R, Martinelli M et al. Prospective validation of ATA and ETA sonographic pattern risk of thyroid nodules selected for FNAC. J Clin Endocrinol Metab 2018;103:2362–2368.