Limitations of the 2015 ATA Guidelines for Prediction of Thyroid Cancer: A Review of 1947 Consecutive Aspirations

Amit Pandya,[1] Elaine M. Caoili,[1] Farah Jawad-Makki,[1] Ashish P. Wasnik,[1] Prasad R. Shankar,[1] Ron Bude,[1] Megan R. Haymart,[2] and Matthew S. Davenport[1]

[1]Department of Radiology, University of Michigan Health Systems, Ann Arbor, Michigan 48109; and [2]Department of Internal Medicine, University of Michigan Health Systems, Ann Arbor, Michigan 48109

Background: The 2015 American Thyroid Association (ATA) guidelines have been proposed to aid in the management of thyroid nodules by determining whether fine needle aspiration is indicated.

Objective: To determine whether the ATA guidelines contribute to the overdiagnosis of thyroid cancer.

Patients and Methods: This was a retrospective cohort study of ultrasound-imaged thyroid nodules (n = 1947) consecutively aspirated at a tertiary care center from 1 October 2009 to 22 February 2016. Nodules were retrospectively reviewed, assigned a 2015 ATA morphology, and placed into one of five 2015 ATA categories of risk (ATA-1, <1% risk of malignancy; ATA-2, <3% risk; ATA-3, 5% to 10% risk, ATA-4: 10% to 20% risk; ATA-5, >70% to 90% risk) by a reader who was blinded to cytology. ATA category was compared with cytopathology. The positive predictive value (PPV) of each ATA category was calculated with respect to cancer. Numbers needed to aspirate and Pearson correlations were calculated. Interrater agreement for ATA category across five readers was assessed.

Results: The PPV for cancer increased by ATA category [category 1 to 5, respectively: 0% (0/14), 2% (4/249), 5% (36/733), 12% (104/850), 28% (28/101)]. The number needed to sample to detect one papillary cancer was 125 (ATA-2), 49 (ATA-3), 13 (ATA-4), and 5 (ATA-5). The overall interrater agreement for ATA score across all five readers was fair (intraclass correlation coefficient 0.460).

Conclusions: The 2015 ATA guidelines stratify risk for thyroid cancer; however, the stratification system is overly optimistic regarding cancer detection rates for the higher-risk nodules, and there is only fair interrater agreement. (***J Clin Endocrinol Metab*** **103: 3496–3502, 2018)**

Thyroid nodules are common and increasingly discovered with the use of imaging, particularly ultrasound. Ultrasound can detect incidental nodules in ≤68% of adults, with a higher prevalence found in older adults and women (1, 2). The evaluation of a thyroid nodule often includes an assessment for malignant potential, which typically includes ultrasound characterization and fine needle aspiration (FNA). Unfortunately, the sonographic appearance of these nodules varies, and no single imaging feature is predictive of malignancy. Thus, risk stratification systems such as the 2015 American Thyroid Association (ATA) guidelines (3) and the American College of Radiology Thyroid Imaging Reporting and Data System (4) have been designed and revised to help physicians treat patients with thyroid nodules.

However, risk stratification is complicated because detection of thyroid cancer has been linked to substantial harm related to the overdiagnosis and overtreatment of low-risk disease. Imaging-based classification systems do

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA

Received 12 April 2018. Accepted 27 June 2018.
First Published Online 3 July 2018

Abbreviations: ATA, American Thyroid Association; FNA, fine needle aspiration; ICC, intraclass correlation coefficient; PPV, positive predictive value.

not clearly differentiate higher-risk cancers from lower-risk cancers (3, 4); rather, they are designed to assess the risk of a nodule being malignant. In those with a reasonable risk of malignancy, FNA is usually recommended, and the histology from that sampling can lead to downstream consequences that may be harmful to the patient (*e.g.,* anxiety, unneeded thyroidectomy, and its associated morbidities). In light of recent evidence challenging routine use of thyroid ultrasound, determining whether imaging-based risk stratification is effective for the detection of clinically significant disease is important. Our purpose was to determine whether the 2015 ATA guidelines effectively stratify thyroid nodules and reduce unnecessary interventions.

Materials and Methods

With institutional review board approval and in compliance with the Health Insurance Portability and Accountability Act, pertinent imaging studies and medical records of the patients in this study were accessed. Participant informed consent was waived by the institutional review board because of the retrospective nature of this investigation.

Subjects

All subjects undergoing first-time FNA of a thyroid nodule in the radiology department at our institution between October 2009 and February 2016 were identified *via* the electronic medical record system and Department of Radiology records. From this group, 28 patients had undergone repeat procedural visits for FNA of a thyroid nodule, and thus only the most recent procedure was included. For patients who underwent FNA of several nodules in one procedural visit, only the first nodule sampled was included in the data. There were no other exclusion criteria. The final study population consisted of 1947 thyroid nodules in 1947 subjects.

Subjects underwent an ultrasound-guided FNA performed by a member of our cross-sectional interventional service. Before the aspiration procedure, subjects underwent a diagnostic thyroid ultrasound to evaluate the targeted nodules. The diagnostic thyroid ultrasound and FNA were performed with electronically focused linear transducers ranging in frequency from 6 to 15 mHz (GE Logiq 700; GE Medical Systems, Milwaukee, WI; and ATL 3000 and 5000; Advanced Technology Laboratories, Bothell, WA). Aspirations were performed with a series of 25-gauge needles and free-hand technique under direct sonographic visualization. The needle was inserted into the targeted nodules, and aspirations were performed with the capillary method. Varying areas of the nodule were sampled in each pass. A minimum of six passes were performed unless a cytopathologist was present. In these latter cases, cellular adequacy was determined by the cytopathologist, and sampling was stopped if cellular adequacy was obtained. The maximum number of passes was 12.

Image review

One radiologist, with 9 years' experience, was blinded to the cytology results. This radiologist retrospectively reviewed the diagnostic thyroid ultrasound images on a picture archiving and communication workstation, determined whether each nodule had microcalcifications, assigned each nodule one of 14 morphologic descriptors according to the 2015 ATA guidelines (3), and placed each nodule into one of five 2015 ATA categories of risk (ATA-1, <1% risk of malignancy; ATA-2, <3% risk; ATA-3, 5% to 10% risk; ATA-4, 10% to 20% risk, ATA-5, >70% to 90% risk). Cysts were placed in ATA-1. Nodules that were interpreted to be spongiform or partially cystic were placed in ATA-2. Nodules that appeared hyperechoic solid with a regular margin, isoechoic with a regular margin, or partially cystic with an eccentric solid area or areas were categorized as ATA-3. Nodules that were hypoechoic solid with a regular margin were placed into ATA-4. Nodules that were hypoechoic with an irregular margin with or without microcalcifications, hypoechoic taller than wide, hypoechoic with an irregular margin and extrathyroidal extension, hypoechoic with interrupted rim calcification with soft tissue extension, or irregular margins with suspicious ipsilateral lymph nodes were placed into ATA-5. Nodule echogenicity, margins, shape, cystic nature, and presence of microcalcifications were assessed according to 2015 ATA guidelines (3).

A subset of 180 nodules from nine image groups (listed below) was then derived from the entire data set to be used for calculating interrater agreement. This subset included all subclassifications of ATA risk that had ≥20 nodules represented as classified by the initial reader. For subclassifications with >20 eligible nodules, a set of 20 was randomly selected from that subclassification for inclusion *via* a random number generator (Microsoft Excel; Microsoft Corporation, Redmond, WA). The following subclassifications were included:

ATA-2: partially cystic without suspicious features (n = 20)
ATA-2: spongiform (n = 20)
ATA-3: hyperechoic with solid regular margins (n = 20)
ATA-3: isoechoic with solid regular margins (n = 20)
ATA-3: partially cystic with eccentric solid area (n = 20)
ATA-3: partially cystic with multiple eccentric solid areas (n = 20)
ATA-4: hypoechoic with solid regular margin (n = 20)
ATA-5: hypoechoic with irregular margins (n= 20)
ATA-5: hypoechoic with irregular margins and microcalcifications (n = 20)

This subset was reviewed by four additional radiologists (with 2 to 30 years of experience) who were blinded to cytology results. These readers retrospectively reviewed the diagnostic thyroid ultrasound images on a picture archiving and communication workstation and determined whether each nodule had microcalcifications, assigned each nodule one of 14 morphologic descriptors according to the 2015 ATA guidelines (3), and placed each nodule into one of five 2015 ATA categories of risk (3).

Cytopathologic reference standard

All thyroid FNAs were interpreted according to the Bethesda system (5) as recommended by the 2015 ATA guidelines (3). The Bethesda system (5) for reporting thyroid cytology is based on the 2007 National Cancer Institute State of Science Conference and provides six diagnostic categories that indicate the risk of cancer. These categories and their risk of malignancy are nondiagnostic, 1% to 4%; benign, 0% to 3%; atypia of undetermined significance, ~5% to 15%; suspicious for follicular neoplasm, 15% to 30%; suspicious for malignancy, 60% to 75%; and malignancy 97% to 99%.

For subjects whose initial FNA results were inconclusive (*i.e.,* nondiagnostic, atypia or follicular lesion of undetermined

significance, or suspicious for neoplasm), we reviewed the electronic medical record to determine whether a subsequent targeted FNA or surgery was performed to enable a more definitive diagnosis within a year of the initial FNA. In such cases, that final diagnosis was recorded. In cases where no definitive diagnosis was obtained, the initial cytopathology was considered the final result.

Data analysis

Data were summarized with descriptive statistics. The likelihood of thyroid cancer was compared across different demographic groups with a χ^2 test for categorical variables and Student t test for continuous variables. The positive predictive value (PPV) of each ATA category was calculated with respect to cancer, cancer or atypia, and any nonpapillary cancer. Final cytology was used to determine the incidence of any nonpapillary cancer. The relationship between ATA category and risk of either any cancer or atypia or any nonpapillary cancer was assessed with Pearson correlation. Numbers needed to sample were calculated for each ATA category.

Interrater agreement for ATA score was assessed with single-measure two-way random-effects intraclass correlation coefficients (ICCs). Interrater agreement for the presence of microcalcifications was assessed with κ statistics (6). The 95% CIs were calculated. Levels of agreement for ICC were assessed as <0.40 (poor), 0.40 to 0.59 (fair), 0.60 to 0.74 (good), and 0.75 to 1.00 (excellent). Levels of agreement for κ were assessed as 0.00 to 0.20 (slight), 0.21 to 0.40 (fair), 0.41 to 0.60 (moderate), 0.61 to 0.80 (substantial), and 0.81 to 1.00 (almost perfect).

Results

Our study population consisted of 1472 women and 475 men with a mean age of 56 years (range, 26 to 86 years). All subjects remained stable throughout the procedure and were discharged the same day. There were no serious complications.

The mean thyroid nodule diameter was 1.7 cm ± 0.9 cm. Most nodules on initial aspiration were benign [64% (1252/1947)] or nondiagnostic [16% (314/1947)] (Tables 1 and 2). The nondiagnostic rate was weakly negatively correlated with ATA classification ($r^2 = -0.21$), indicating that higher ATA scores were modestly less likely to be nondiagnostic. A minority [12% (226/1947)] of subjects had both an indeterminate initial cytopathology and an indeterminate final diagnosis on follow-up.

Initial results were suspicious for, or diagnostic of, cancer in 10% (193/1947) and diagnostic of nonpapillary cancer in 1% (19/1947). In addition to papillary thyroid cancer, other malignancies included follicular thyroid cancer, medullary carcinoma, and Hurthle cell carcinoma as well as metastases, poorly differentiated carcinoma, leukemia, lymphoma, and squamous cell carcinoma. Nonpapillary cancers were significantly larger than other nodules (mean diameter, 2.7 ± 1.6 cm; $P = 0.01$; Table 1), and thyroid nodules in men were more likely to be malignant [10% (48/475) vs 3% (50/1472); $P < 0.0001$] (Table 1).

The PPV for cancer strongly correlated with ATA category, but the PPV for greater ATA scores was less than is stated in the 2015 ATA guidelines (3): ATA-1, 0% (0/14); ATA-2, 2% (4/249); ATA-3, 5% (36/733); ATA-4, 12% (104/850); and ATA-5, 28% (28/101). The PPV for cancer or atypia was also strongly correlated with ATA category, but the PPV for greater ATA scores was also less than is stated in the 2015 ATA guidelines (3): ATA-1, 0% (0/14); ATA-2, 6% (15/249); ATA-3, 18% (133/733); ATA-4, 24% (201/850); and ATA-5, 32% (32/101). The strong correlation between ATA score and likelihood of malignancy was similar for nonpapillary cancers, but the risk of nonpapillary cancer was an order of magnitude less than the risk of any cancer or atypia: ATA-1, 0% (0/14); ATA-2, 0.8% (2/249); ATA-3, 2.9% (21/733); ATA-4, 4.8% (41/850); and ATA-5, 5.9% (6/101). The number needed to sample to detect one cancer was 62 (ATA-2), 20 (ATA-3), 8 (ATA-4), and 4 (ATA-5). The number needed to sample to detect one papillary cancer was 125 (ATA-2), 49 (ATA-3), 13 (ATA-4), and 5 (ATA-5).The number needed to biopsy to detect one nonpapillary cancer was 125 (ATA-2), 35 (ATA-3), 21 (ATA-4), and 17 (ATA-5), which was an order of magnitude higher than for any cancer or atypia: 17 (ATA-2), 6 (ATA-3), 4 (ATA-4), and 3 (ATA-5).

Nodules with an initial cytology result of nondiagnostic (n = 314) were found to be malignant in only

Table 1. Relationship of Demographic Details to Cytopathology Outcome

		Cytology on FNA						
Parameter	Overall	Nondiagnostic	Benign	Atypia or Follicular Lesion; Uncertain Significance	Suspicious for Follicular Neoplasm	Suspicious for Malignancy	Malignancy: Papillary Cancer	Malignancy: Nonpapillary Cancer
All subjects	1947	314 (16%)	1252 (64%)	209 (11%)	64 (3%)	31 (2%)	79 (4%)	19 (1%)
Female	1472	237 (16%)	987 (67%)	148 (10%)	50 (3%)	21 (1%)	40 (3%)	10 (1%)
Male	475	77 (16%)	265 (56%)	61 (13%)	14 (3%)	10 (2%)	39 (8%)	9 (2%)
Age, y	56 ± 15	56 ± 14	56 ± 15	57 ± 14	54 ± 17	46 ± 15	51 ± 15	69 ± 9
Mean nodule diameter, cm	1.68 ± 0.88	1.67 ± 0.86	1.64 ± 0.83	1.83 ± 0.96	2.06 ± 1.05	1.44 ± 0.77	1.64 ± 0.97	2.74 ± 1.60
Nodule volume, mL	9.9 ± 23.7	9.2 ± 17.9	8.9 ± 23.0	12.5 ± 26.5	16.9 ± 30.9	6.0 ± 9.8	10.6 ± 26.4	43.1 ± 60.2

Denominators for percentages are the total number of subjects in each row.

Table 2. Prevalence and Initial Cytopathology Outcomes of 1947 Thyroid Nodules Sampled From 1 October 2009 to 22 February 2016, Stratified by the 2015 ATA Classification System and the ATA-Assigned Purported Risk of Malignancy

		Cytology on FNA						
ATA Nodule Type	N	Nondiagnostic	Benign	Atypia or Follicular Lesion; Uncertain Significance	Suspicious for Follicular Neoplasm	Suspicious for Malignancy	Malignancy: Papillary Cancer	Malignancy: Nonpapillary Cancer
Class 1 (<1% malignant)	14	6 (43%)	8 (57%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Cyst	14	6 (43%)	8 (57%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Class 2 (<3% malignant)	249	34 (14%)	200 (80%)	11 (4%)	1 (0.4%)	1 (0.4%)	2 (1%)	0 (0%)
Spongiform	119	13 (11%)	99 (83%)	6 (5%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)
Partially cystic, no suspicious features	130	21 (16%)	101 (78%)	5 (4%)	1 (0.8%)	1 (0.8%)	1 (0.8%)	0 (0%)
Class 3 (5%–10% malignant)	733	102 (14%)	498 (68%)	97 (13%)	18 (2%)	7 (1%)	10 (1%)	1 (0.1%)
Hyperechoic, solid regular margin	355	53 (15%)	229 (65%)	50 (14%)	13 (4%)	5 (1%)	4 (1%)	1 (0.3%)
Isoechoic, solid regular margin	207	29 (14%)	136 (66%)	34 (16%)	3 (1%)	1 (0.5%)	4 (2%)	0 (0%)
Partially cystic, eccentric solid area	124	17 (14%)	96 (77%)	6 (5%)	2 (2%)	1 (0.8%)	2 (2%)	0 (0%)
Partially cystic, >1 eccentric solid areas	47	3 (6%)	37 (79%)	7 (15%)	0 (0%)	0 (%)	0 (0%)	0 (0%)
Class 4 (10%–20% malignant)	850	162 (19%)	487 (57%)	97 (11%)	23 (3%)	21 (3%)	47 (6%)	13 (2%)
Hypoechoic, solid regular margin	850	162 (19%)	487 (57%)	97 (11%)	23 (3%)	21 (3%)	47 (6%)	13 (2%)
Class 5 (>70%–90% malignant)	101	10 (10%)	59 (58%)	4 (4%)	1 (1%)	2 (2%)	20 (20%)	5 (5%)
Microcalcifications, hypoechoic, irregular margin	52	5 (10%)	27 (52%)	1 (2%)	1 (2%)	2 (4%)	16 (31%)	0 (0%)
Hypoechoic, irregular margin	48	5 (10%)	31 (65%)	3 (6%)	0 (0%)	0 (0%)	4 (8%)	5 (10%)
Hypoechoic, taller than wide	1	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Hypoechoic, irregular margin, extrathyroidal extension	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hypoechoic, interrupted rim calcification, soft tissue extrusion	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Irregular margins, suspicious node	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Denominators for percentages are the number of nodules in each row.

Abbreviation: N/A, not applicable.

2% (5/314) on follow-up, including three nonmicroscopic papillary cancers and two nodules suspicious for follicular cancer (Table 3). Nodules initially considered suspicious for malignancy had a high likelihood of a final malignant diagnosis on follow-up, ranging from 49% (18/43) for nodules originally considered suspicious for follicular neoplasm to 84% (26/31) for nodules originally considered suspicious for malignancy (Table 3). The histologic finding of microscopic papillary thyroid carcinoma (n = 12) was identified incidentally in the setting of a benign hyperplasia and was not classified as papillary carcinoma or malignancy.

The overall interrater agreement for ATA score across all five readers was fair (ICC, 0.460; 95% CI, 0.391–0.533), with pairwise agreement ranging from 0.366–0.606 (ICC). Pairwise interrater agreement for presence of microcalcifications was slight to fair (pairwise κ range 0.029–0.400) (Table 4).

Discussion

In the United States and many other countries, the incidence of thyroid cancer has increased steadily. In the past decade, the incidence of this cancer has increased on average by 3.8% each year in the United States (7–9). If these trends continue, thyroid cancer may move from the 11th most common cancer in the United States to the 4th most common cancer by 2030 (7, 8). This increased incidence is caused by increased diagnosis of occult cancers and subclinical lesions through a number of pathways, including thyroidectomy for benign disease and diagnostic imaging such as CT and ultrasound (8–11). Nearly all newly diagnosed thyroid cancers are differentiated slow-growing papillary cancers in nodules ≤1 cm in size. Detection of aggressive disease has remained unchanged, as has thyroid cancer mortality (7–9). however, the number of thyroidectomies continues to increase (9).

As an example, the South Korean government in 1999 began a national screening program that allowed inexpensive ultrasound imaging of the neck for thyroid cancer. This program resulted in a 15-fold increase in the detection of thyroid cancer, which surpassed the detection of breast cancer and other common malignancies. Nearly all patients were treated with either radical (67%)

Table 3. Final Outcomes of 597 Thyroid Nodules That Had Initial FNA Results of Nondiagnostic (n = 314), Atypia or Follicular Lesion of Undetermined Significance (n =209), Suspicious for Follicular Neoplasm (n = 43), or Suspicious for Malignancy (n = 31)

		Cytology on FNA			
Final Diagnosis	**N**	**Nondiagnostic**	**Atypia or Follicular Lesion; Uncertain Significance**	**Suspicious for Follicular Neoplasm**	**Suspicious for Malignancy**
All nodules	597	314	209	43	31
Benign findings	295	50% (n = 158)	57% (n = 120)	56% (n = 24)	16% (n = 5)
Colloid nodule	30	13	12	4	1
No malignant cells	12	5	7	0	0
Nodular hyperplasia	218	124	78	12	4
Probably benign	24	6	10	8	0
Thyroiditis	23	10	13	0	0
Indeterminate findings	226	48% (n = 151)	35% (n = 74)	2% (n = 1)	0% (n = 0)
Atypical cells	3	1	2	0	0
Atypical lymphoid infiltrate	1	0	1	0	0
Follicular lesion of undetermined significance	69	9	60	0	0
Hurthle cell of undetermined significance	12	2	9	1	0
Nondiagnostic	141	139	2	0	0
Malignant findings	76	2% (n = 5)	7% (n = 15)	42% (n = 18)	84% (n = 26)
Suspicious for malignancy	2	0	0	0	2
Papillary cancer, suspicious for	1	0	0	1	0
Papillary cancer	45	3	13	8	21
Follicular cancer, low-grade	2	0	0	2	0
Follicular cancer, suspicious for	11	2	2	7	0
Hurthle cell carcinoma	1	0	0	0	1
Medullary carcinoma	2	0	0	0	2

Denominators are the total number of nodules in each column.

or subtotal thyroidectomy (33%). Insurance claim analysis of patients who underwent thyroid surgery in this region revealed complications such as hypoparathyroidism (11%), vocal cord paralysis (2%), and potentially lifelong thyroid replacement therapy (12). As a result of such data, the US Preventive Services Task Force has recommended against screening for thyroid cancer in asymptomatic adults (13).

In the most recent version of the ATA guidelines (2015), the ATA stresses minimizing harm from the overtreatment of low-risk patients while providing appropriate treatment of those at higher risk (3). However, it is unclear whether the ATA risk stratification system was optimized for detection of any cancer (including indolent cancers for which early detection is unlikely to benefit the patient) or for higher-risk cancer more specifically, especially when the risk estimates are greatly increased with higher ATA categories. Our data demonstrate that the ATA system does effectively stratify risk of thyroid cancer but with the indirect harm of detecting an order of magnitude more low-risk disease (7). Based on our data, with the ATA scores used as a threshold to stratify the need for thyroid nodule sampling, the number needed to sample or aspirate to detect cancer was 62 (ATA-2), 20 (ATA-3), 8 (ATA-4), and 4 (ATA-5), which was higher than for detecting either cancer or atypia: 17 (ATA-2), 6 (ATA-3), 4 (ATA-4), and 3 (ATA-5).

The PPVs we observed for ATA-3, ATA-4, and ATA-5 observations were less than or in the lower range reported in the ATA guidelines. For example, the stated risk of malignancy in an ATA-3 nodule is reported to be 5% to 10% (3), and we observed a rate of 5%. The stated risk in an ATA-4 nodule is 10% to 20%, and we observed a rate of 12%. The stated risk in an ATA-5 nodule is >70% to 90%, and we observed a rate of 28%, which is far less than what is predicted by the ATA classification. This implies that the ATA guidelines are overly optimistic regarding the cancer detection rates for higher-risk ATA nodules. According to the ATA guidelines, highly suspicious nodules detected on ultrasound should undergo repeat imaging and FNA within 12 months; however, in our study group, repeat FNA may not be practical with the low PPV associated with ATA-5 nodules. Once a nodule is detected, active surveillance may be a promising strategy in patients with both low-risk disease and high-risk disease categories based on the actual observed rate of malignancy found in this study. Active surveillance has been evaluated in prospective trials by Ito *et al.* (14, 15). These investigators studied a cohort of patients who underwent immediate surgery

Table 4. Pairwise Interrater Agreement for ATA Scoring (Scale: 2–5) and Presence of Microcalcifications Between 5 Blinded Independent Readers

Reader and Outcome	Reader 2	Reader 3	Reader 4	Reader 5
ATA score				
Reader 1	0.414 (0.285–0.528)	0.424 (0.296–0.537)	0.506 (0.388–0.607)	0.396 (0.265–0.513)
Reader 2	—	0.445 (0.320–0.555)	0.490 (0.370–0.593)	0.412 (0.283–0.526)
Reader 3	—	—	0.522 (0.407–0.621)	0.606 (0.505–0.691)
Reader 4	—	—	—	0.366 (0.232–0.486)
Microcalcifications				
Reader 1	0.335 (0.200–0.470)	0.400 (0.273–0.527)	0.212 (0.098–0.326)	0.029 (−0.034–0.092)
Reader 2	—	0.321 (0.156–0.486)	0.200 (0.037–0.363)	0.042 (−0.062–0.146)
Reader 3	—	—	0.204 (0.030–0.378)	0.111 (−0.028–0.250)
Reader 4	—	—	—	0.045 (−0.114–0.204)

Data for ATA scores are expressed with single-measure two-way random effects intraclass correlation statistics, and data for presence of microcalcifications are expressed with κ statistics; numbers in parentheses are 95% CIs. The overall interrater agreement for ATA score across all five readers was fair (ICC, 0.460; 95% CI, 0.391–0.533).

after the diagnosis of papillary thyroid cancer, whereas another cohort underwent surveillance. After 4 to 6 years of follow-up, there was no difference in mortality rates between the two groups; no deaths occurred, and no subjects on active surveillance developed metastatic disease or locoregional spread. Such trials highlight the need for improved management strategies in patients with thyroid nodules with a low risk of malignancy and those with suspected or known thyroid cancer. These trials also demonstrate the need for more rigorous classifications to guide management.

Our study has limitations. Although we included a large number of consecutive subjects undergoing first-time FNA of thyroid nodules over a long period of time, because of the retrospective design it is possible that selection biases at our institution could affect the PPVs of the ATA categories we analyzed. Although 1947 nodules is a large sample size, it may not reflect the higher-risk populations seen by endocrinologists, where the prevalence of thyroid malignancy may be higher. It is hypothesized that the ATA classification may be more appropriate in this setting. Another limitation is that the primary ATA classification was assigned by one experienced expert in ultrasound who is familiar with the ATA guidelines. That individual's performance is confirmed by the strong relationship between assigned ATA scores and risk of malignancy, but it is possible the results would have been different if other raters had evaluated the entire group of nodules. We attempted to account for this limitation by measuring interrater agreement across five readers in a subset of 180 nodules. We found that the reliability or agreement among the readers was only fair. Therefore, on a per-patient basis, the ATA scores are likely to vary between different readers. The lack of reproducibility of results in our practice probably reflects the lack of reproducibility in multiple practices throughout the country. Although this is an acknowledged limitation, this study may be a more accurate representation of applying the criteria across a variety of practitioners interpreting thyroid ultrasound. Another limitation is that a majority of nodules lacked histological confirmation of the initial FNA, indicating that a definitive diagnosis was unknown. Within this group, 226 subjects (12%) had both an indeterminate initial cytology and an indeterminate final diagnosis. This limitation is similar to more recent publications that note that varying rates of malignancy have been reported in indeterminate nodules and thus demonstrates the need for a large series evaluating this group (16). Fortunately, the Bethesda system for reporting thyroid cytopathology has recently been revised, incorporating new data and developments in the management of thyroid nodules. The 2017 system has substantiated, strengthened, and improved the estimation of risk of malignancy associated with the six original categories: nondiagnostic, 5% to 10%; benign, 0% to 3%; atypia of undetermined significance or follicular lesion of undetermined significance, 10% to 30%; follicular neoplasm or suspicious for a follicular neoplasm, 25% to 40%; suspicious for malignancy, 50% to 75%; and malignant, 97% to 99% (17). These risks are similar to those reported with the previous Bethesda system. Finally, the impact of molecular diagnostics was not included in this study because this was not the primary objective of our study. Molecular diagnostics is an important component in today's current practice. However its impact is still in question. Dedhia *et al.* (18) reported that in their practice, molecular diagnostics had limited effect in preventing thyroidectomies in patients found to have indeterminate nodules.

In conclusion, the 2015 ATA guidelines effectively stratify risk for cancer, but the predictive value for detecting thyroid cancer is low, particularly in nodules considered to have higher risk of malignancy. Detection of these cancers is thus accompanied by the harm of detecting an order of magnitude more low-risk disease and can result in the overtreatment of a large number of patients. Improved

methods of risk stratification are needed to identify those who would benefit most from immediate FNA and ultimately surgery. In addition, the effectiveness of the guidelines may be hampered by varied interpretations of the sonographic classification of thyroid nodules.

Acknowledgments

Correspondence and Reprint Requests: Elaine M. Caoili, MD, Department of Radiology, University of Michigan Health Systems, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109. E-mail: caoili@med.umich.edu.

Disclosure Summary: The authors have nothing to disclose.

References

1. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest*. 2009;**39**(8):699–706.
2. Ezzat S, Sarti DA, Cain DR, Braunstein GD. Thyroid incidentalomas. Prevalence by palpation and ultrasonography. *Arch Intern Med*. 1994;**154**(16):1838–1840.
3. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 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):1–133.
4. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, Stavros AT. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR-TI-RADS committee. *J Appl Commun Res*. 2017;**14**:587–595.
5. Cibas ES, Ali SZ; NCI Thyroid FNA State of the Science Conference. The Bethesda system for reporting thyroid cytopathology. *Am J Clin Pathol*. 2009;**132**(5):658–665.
6. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;**8**(1):23–34.
7. Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (Eds.). SEER cancer statistics review, 1975–2014. Bethesda, MD: National Cancer Institute. Available at: www.seer.cancer.gov/csr/1975_2014/. Accessed May 2018.
8. La Vecchia C, Malvezzi M, Bosetti C, Garavello W, Bertuccio P, Levi F, Negri E. Thyroid cancer mortality and incidence: a global overview. *Int J Cancer*. 2015;**136**(9):2187–2195.
9. Brito JP, Al Nofal A, Montori VM, Hay ID, Morris JC. The impact of subclinical disease and mechanism of detection on the rise in thyroid cancer incidence: a population-based study in Olmstead County, Minnesota during 1935 through 2012. *Thyroid*. 2015;**25**(9):999–1007.
10. Leenhardt L, Bernier MO, Boin-Pineau MH, Conte Devolx B, Maréchaud R, Niccoli-Sire P, Nocaudie M, Orgiazzi J, Schlumberger M, Wémeau JL, Chérie-Challine L, De Vathaire F. Advances in diagnostic practices affect thyroid cancer incidence in France. *Eur J Endocrinol*. 2004;**150**(2):133–139.
11. Davies L, Welch HG. Current thyroid cancer trends in the United States. *JAMA Otolaryngol Head Neck Surg*. 2014;**140**(4):317–322.
12. Ahn HS, Kim HJ, Welch HG. Korea's thyroid-cancer "epidemic"–screening and overdiagnosis. *N Engl J Med*. 2014;**371**(19):1765–1767.
13. Bibbins-Domingo K, Grossman DC, Curry SJ, Barry MJ, Davidson KW, Doubeni CA, Epling JW Jr, Kemper AR, Krist AH, Kurth AE, Landefeld CS, Mangione CM, Phipps MG, Silverstein M, Simon MA, Siu AL, Tseng CW; US Preventive Services Task Force. Screening for thyroid cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2017;**317**(18):1882–1887.
14. Ito Y, Miyauchi A, Inoue H, Fukushima M, Kihara M, Higashiyama T, Tomoda C, Takamura Y, Kobayashi K, Miya A. An observational trial for papillary thyroid microcarcinoma in Japanese patients. *World J Surg*. 2010;**34**(1):28–35.
15. Ito Y, Miyauchi A. Nonoperative management of low-risk differentiated thyroid carcinoma. *Curr Opin Oncol*. 2015;**27**(1):15–20.
16. Maino F, Forleo R, Martinelli M, Fralassi N, Barbato F, Pilli T, Capezzone M, Brilli L, Ciuoli C, Di Cairano G, Nigi L, Pacini F, Castagna MG. Prospective validation of ATA and ETA sonographic pattern risk of thyroid nodules selected for FNAC. *J Clin Endocrinol Metab*. 2018;**103**(6):2362–2368.
17. Cibas ES, Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. *Thyroid*. 2017;**27**(11):1341–1346.
18. Dedhia PH, Rubio GA, Cohen MS, Miller BS, Gauger PG, Hughes DT. Potential effects of molecular testing of indeterminate thyroid nodule fine needle aspiration biopsy on thyroidectomy volume. *World J Surg*. 2014;**38**(3):634–638.