Quantitative Framework for Risk Stratification of Thyroid Nodules With Ultrasound: A Step Toward Automated Triage of Thyroid Cancer

Alfiia Galimzianova[1]
Sean M. Siebert[2]
Aya Kamaya[2]
Daniel L. Rubin[1]
Terry S. Desser[2]

OBJECTIVE. The purpose of this study was to explore whether a quantitative framework can be used to sonographically differentiate benign and malignant thyroid nodules at a level comparable to that of experts.

MATERIALS AND METHODS. A dataset of ultrasound images of 92 biopsy-confirmed nodules was collected retrospectively. The nodules were delineated and annotated by two expert radiologists using the standardized Thyroid Imaging Reporting and Data System lexicon of the American College of Radiology. In the framework studied, quantitative features of echogenicity, texture, edge sharpness, and margin curvature properties of thyroid nodules were analyzed in a regularized logistic regression model to predict malignancy of a nodule. The framework was validated by leave-one-out cross-validation technique, and ROC AUC, sensitivity, and specificity were used to compare with those obtained with six expert annotation-based classifiers.

RESULTS. The AUC of the proposed method was 0.828 (95% CI, 0.715–0.942), which was greater than or comparable to that of the expert classifiers, for which the AUC values ranged from 0.299 to 0.829 ($p = 0.99$). Use of the proposed framework could have avoided biopsy of 20 of 46 benign nodules in a curative strategy (at sensitivity of 1, statistically significantly higher than three expert classifiers) or helped identify 10 of 46 malignancies in a conservative strategy (at specificity of 1, statistically significantly higher than five expert classifiers).

CONCLUSION. When the proposed quantitative framework was used, thyroid nodule malignancy was predicted at the level of expert classifiers. Such a framework may ultimately prove useful as the basis for a fully automated system of thyroid nodule triage.

Keywords: computer-aided diagnosis, quantitative image analysis, thyroid cancer

doi.org/10.2214/AJR.19.21350

D. L. Rubin and T. S. Desser contributed equally to this study.

Received February 21, 2019; accepted after revision September 25, 2019.

Based on a presentation at the 2017 American Medical Informatics Association annual meeting, Washington, DC.

Supported in part by grants from GE Healthcare (Blue Sky Initiative at Stanford University) and the National Cancer Institute, National Institutes of Health (U01CA142555, 1U01CA190214, and 1U01CA187947).

[1]Department of Biomedical Data Science, Stanford University School of Medicine, 1265 Welch Rd, Palo Alto, CA 94305. Address correspondence to A. Galimzianova (alfiia.galimzianova@gmail.com).

[2]Department of Radiology, Stanford University, Palo Alto, CA.

AJR 2020; 214:885–892

0361–803X/20/2144–885

Detection of both benign and malignant thyroid nodules has increased in the United States. Much of this increase is believed to be due to increased use of imaging [1] with a concomitant increase in detection of asymptomatic thyroid nodules and indolent thyroid malignancies that would likely never have become clinically manifest. Autopsy studies have shown that as many as 50–67% of adults nationwide are expected to have thyroid nodules [2, 3] but that only 0.2% of the population is reported to have thyroid cancer [4].

Current definitive diagnosis of thyroid nodules requires tissue biopsy or even surgery, but only 5–7% of these nodules are found to be malignant [5]. Thus, most patients are inevitably exposed to unnecessary health risks associated with these invasive tests, and societal health care costs increase substantially. There is a critical need for methods to reliably estimate the malignancy risk of thyroid nodules to decrease the number of invasive interventions performed on low-risk benign nodules. Intensive research has been underway in radiology, endocrinology, and surgery to attempt to identify patients at high risk of aggressive thyroid malignancy and who need invasive diagnostic intervention. Despite these efforts, accurate diagnosis based on imaging findings remains challenging.

Ultrasound (US) is the standard of care imaging modality for visually assessing the risks of thyroid nodule malignancy, determining the need for biopsy, and guiding fine-needle aspiration biopsy. Several expert committees have proposed guidelines for triage and management of thyroid nodules detected at US that are being used or incorporated into clinical routine by institutions worldwide. In the United States, the widely used systems include those of the American Thyroid Association (ATA) published in 2015 [6] and the American College of

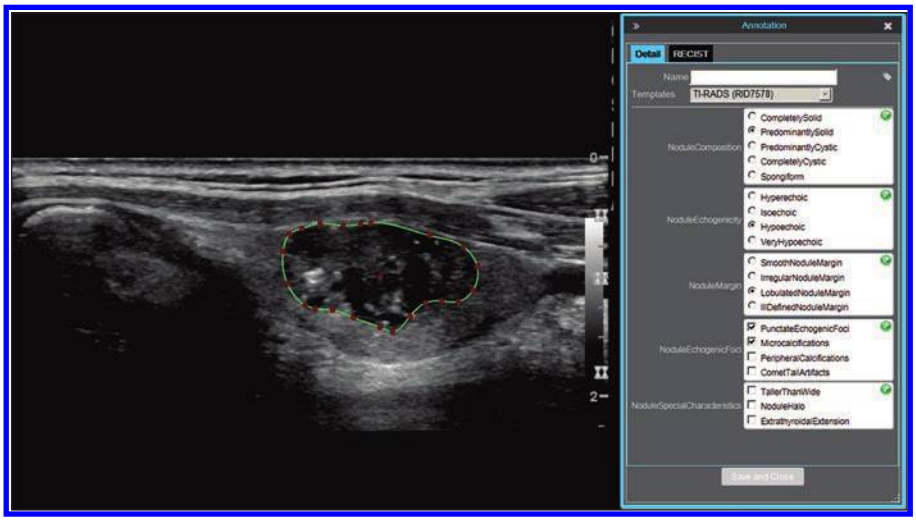

Fig. 1—73-year-old man with papillary carcinoma of left lobe of thyroid. Screen shot shows example of thyroid nodule annotation (segmentation and American College of Radiology Thyroid Imaging Reporting and Data System [TI-RADS] annotation) performed on ultrasound image in longitudinal projection with electronic Physician Annotation Device software (Stanford Medicine Radiology). Radiologists performed nodule segmentation by selecting points (*red*) on nodule outline (*green*) while controlling smoothing of outline polygon by means of spline interpolation. RECIST = Response Evaluation Criteria in Solid Tumors, RID = Radiological Society of North America RadLex identifier.

Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) proposed in 2017 [7]. Other guidelines include those of the British Thyroid Association (BTA) published in 2014 [8], Korean TIRADS (K-TIRADS) [9], and European Thyroid Association TIRADS (EU-TIRADS) [10] that are or are being adopted in the United Kingdom, South Korea, and the rest of Europe.

Among the systems, ACR TI-RADS relies on computing the malignancy score with a set of US features to arrive at a biopsy recommendation. The other systems are based on various nodule appearance patterns associated with varied risks of malignancy. The use of such systems for biopsy decision making helps to standardize recommendations among readers and to reduce the number of unnecessary biopsies. However, explicit risk scoring of each nodule imposes additional burden on radiologists and is subject to intrarater and interrater variability, even among experts [11]. Moreover, the systems used to evaluate nodule malignancy on the basis of analysis of sonographic features are often oversimplified because they are of necessity limited to features appreciable by visual inspection.

By contrast, quantitative analysis of nodule properties by extraction of computational features from US images and estimation of the risks by use of machine learning-based classifiers may help reduce expert labor and eliminate the associated interrater variability. Development of an automated image analysis system would be the first step toward implementation of large-scale automated screening of thyroid image datasets. We propose a quantitative framework for automated cancer risk estimation from US images. In this framework, a rich set of quantitative features and an elastic net classifier are used to estimate the probability that a nodule is cancerous. We cross-validated the framework in a dataset of US images of patients with biopsy-proven diagnoses and compared the findings with those obtained with six expert classification systems.

Materials and Methods

Datasets

US images of patients who underwent US-guided thyroid nodule biopsy at our institution from 2010 to 2015 were collected as part of an institutional review board–approved, HIPAA-compliant retrospective study. In 1181 identified cases, 765 nodules had a longest measured dimension of 1–3 cm; 80 (10.4%) of these nodules were found to be papillary carcinoma. For the purposes of this analysis, 34 patients with malignancy were excluded because they had undergone imaging with an older-generation US system at a frequency less than 12 MHz. The final dataset consisted of 46 malignant and 46 randomly selected size-matched benign (Bethesda category II) nodules from 92 patients (73 women, 19 men; age range, 21–83 years; mean age, 53.8 years) imaged with either a Logiq E9 (GE Healthcare) or an Acuson S2000 (Siemens Healthineers) system.

Annotation

The principal transverse and longitudinal US images of each nodule as selected by a US technician and verified by a radiologist at the examination were included. Two board-certified radiologists (25 and 12 years' experience) specialized in evaluating sonographic studies of thyroid nodules reviewed the images blinded to the diagnosis. The radiologists reviewed the manual annotations and recorded the visual descriptors of nodules specified by the ACR TI-RADS lexicon [12] using electronic Physician Annotation Device software (version 2.8, Stanford Medicine Radiology) [13]. They performed nodule segmentation by selecting points on a nodule outline while controlling the smoothing of the outline polygon by means of spline interpolation (Fig. 1).

Quantitative Features for Automated Malignancy Risk Estimation

To capture the appearance of nodules in a quantitative manner, we selected a rich set of quantitative features to comprehensively encode the echogenicity, texture, edge sharpness, and margin curvature properties of thyroid nodules (Table 1). Nodule echogenicity was expressed by nodule intensity features, that is, first-order intensity statistics and intensity difference between the nodule, its edge, and the surrounding tissue. Nodule texture was expressed as Haralick-based [14] and gray-level cooccurrence matrix [15] features. Nodule edge sharpness was computationally characterized by edge sharpness features comprising statistics of sigmoid fit parameters over the discretized nodule boundary. Nodule shape was characterized by local area integral invariant descriptor [16] statistics at five spatial scales. Overall, 219 computational features were collected for each nodule from both longitudinal and transaxial images in a cumulative manner.

Semantic Ultrasound Features for Radiologist-Based Malignancy Risk Estimation

The nodule descriptors provided by the raters were represented in further analysis as a set of 20 binary features reflecting the presence or absence of the following visual features: composition (solid, predominantly solid, predominantly cystic, cystic, spongiform), echogenicity (hyperechoic, isoechoic, hypoechoic, very hypoechoic), shape (taller than wide), margins (smooth, irregular, lobulated or ill-defined border; complete

TABLE 1: Quantitative Ultrasound Features Used in Proposed Framework

Category	Features
Echogenicity	First-order statistics (energy, entropy, kurtosis, skewness, maximum, 99th percentile, minimum, 1st percentile, mean, median, SD, median absolute deviation, range, range from 1st to 99th percentiles, root mean square, uniformity, variance, 32-bin histogram, deciles, peak) of image intensity, contrast with surrounding tissue, contrast on nodule edge
Texture	Haralick features (autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, variance, sum average, maximum probability, sum variance, sum entropy, difference variance, difference entropy), gray-level run length matrix features (SRE, LRE, GLN, RLN, RP, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE)
Edge sharpness	Sigmoid fit statistics (minimum, maximum, median, mean, median absolute deviation, SD, skewness, kurtosis, 32-bin histogram, deciles) for window and scale parameters
Margin curvature	Statistics (mean, SD, maximum, skewness) and Haar transform of local area integral invariant shape descriptor at five scales

Note—SRE = short-run emphasis, LRE = long-run emphasis, GLN = gray-level nonuniformity, RLN = run-length nonuniformity, RP = run percentage, LGRE = low gray-level run emphasis, HGRE = high gray-level run emphasis, SRLGE = short-run low-gray-level emphasis, SRHGE = short-run high-gray-level emphasis, LRLGE = long-run low-gray-level emphasis, LRHGE = long-run high-gray-level emphasis.

halo; extrathyroidal extension), and echogenic foci (punctate echogenic foci, macrocalcifications, peripheral calcifications, comet-tail artifacts). For reference malignancy risk estimators, we used six expert annotation-based classification systems. Using the ACR TI-RADS lexicon, we reformulated and implemented five guidelines—ATA 2015 [6], ACR TI-RADS 2017 [7], BTA 2014 [8], K-TIRADS 2016 [9], and EU-TIRADS 2017 [10]—and a system described in the work of Smith-Bindman et al. [17]. The semantic features by classifiers are summarized in Table 2.

Evaluation

The quantitative features were used to train the elastic net classifier [18] on the biopsy-confirmed diagnoses of the nodules in two-label formulation, that is, benign versus malignant. Elastic net classification entails a regression model trained with L_1

TABLE 2: Semantic Ultrasound Features Derived From American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) Lexicon and Used in the Classification Systems

Ultrasound Features	Expert Classifiers					
Binary descriptor	ACR TI-RADS 2017	Smith-Bindman et al.	EU-TIRADS 2017	K-TIRADS 2016	ATA 2015	BTA 2014
Composition						
Solid	X	X		X	X	X
Predominantly solid	X			X	X	
Predominantly cystic	X			X	X	
Cystic		X	X	X	X	
Spongiform			X	X	X	X
Echogenicity						
Hyperechoic	X		X	X	X	X
Isoechoic	X		X	X	X	X
Hypoechoic	X		X	X	X	X
Very hypoechoic	X		X	X	X	X
Margins						
Smooth			X			
Irregular	X		X	X		X
Lobulated	X			X		X
Ill defined	X					
Halo						X
Extrathyroidal extension	X					
Shape						
Taller than wide	X			X		X
Echogenic foci						
Punctate echogenic foci	X	X	X	X	X	X
Macrocalcifications	X					X
Peripheral calcifications	X				X	X
Comet-tail artifacts				X		X

Note—The classification of Smith-Bindman et al. can be found in [17]. EU-TIRADS = European Thyroid Association TIRADS, K-TIRADS = Korean TIRADS, ATA = American Thyroid Association, BTA = British Thyroid Association, X = feature used in classifier.

TABLE 3: Annotation Agreement of the American College of Radiology Thyroid Imaging Reporting and Data System Ultrasound Features for Two Raters

Ultrasound Feature	Agreement (%)
Solid	91.3
Predominantly solid	88.0
Predominantly cystic	98.9
Cystic	100
Spongiform	97.8
Hyperechoic	96.7
Isoechoic	84.8
Hypoechoic	73.9
Very hypoechoic	85.9
Smooth	84.8
Irregular	87.0
Lobulated	79.3
Ill-defined	88.0
Halo	85.9
Extrathyroidal extension	97.8
Taller-than-wide	85.9
Punctate echogenic foci	81.5
Macrocalcifications	87.0
Peripheral calcifications	95.7
Comet-tail artifacts	97.8

and L_2 norm regularization of the weights and is particularly beneficial when the number of features is greater than the number of observations [18].

Leave-one-out cross-validation (LOOCV) was performed [19]. Under the LOOCV scheme, n experiments (where n is the sample size) are conducted with each having a unique split of $n - 1$ training and one held out testing observations. The statistics of the resulting n-value vector of independently inferred predictions is used for analysis of performance. By its design, LOOCV ensures that the same sample is never included in the test and train subset of such experiment. LOOCV was used as the evaluation method because of the small sample size, which precluded having separate training and test sets. On each of the LOOCV folds, model selection was performed by threefold cross validation of the training set to choose the model with the lowest mean cross-validated error [20].

For the semantic features, we used as reference malignancy risk estimators three international TIRADS systems (ACR TI-RADS, K-TIRADS, and EU-TIRADS), two guidelines (ATA and BTA), and the proposal by Smith-Bindman et al. [17] designed to lower the rate of unnecessary biopsies. The annotations provided by the two radiologists were transformed into a mineable feature set consisting of 20 binary descriptors used by the systems to score the malignancy risk or find a corresponding pattern (Table 2).

We compared the scoring performance of the implemented scoring systems using ROC curves and AUC. We computed the point and interval estimates (95% CI) using the Mann-Whitney statistic and evaluated the difference between the AUC values using the paired DeLong test [21] at 0.05 significance level. The diagnostic performance of the biopsy recommendations of the stratification systems was evaluated by means of sensitivity and specificity measures with corresponding Clopper-Pearson CIs (at 95%). Improvement in these measures was evaluated for all the pairs of systems by use of the one-tailed McNemar exact test at 0.05 significance level. For all comparisons, annotations from both raters were used independently. Agreement between the raters' annotations was assessed with the Cohen kappa coefficient.

Results

The agreement of semantic feature annotation between the two radiologists ranged between 73.9% and 100% (Table 3). The LOOCV experiment had a mean training time of 33.9 ± 3.4 (SD) seconds and test time of $5.9 \times 10^{-6} \pm 5.9 \times 10^{-7}$ seconds.

Scoring Performance

The ROC curves for the scoring performance of each of the six expert classification systems versus the proposed quantitative framework are shown in Figure 2. The highest performance values were achieved by ACR TI-RADS (AUC, 0.829; 95% CI, 0.726–0.932) and the two raters (AUC, 0.826; 95% CI, 0.725–0.926). The quantitative framework achieved an AUC of 0.828 (95%

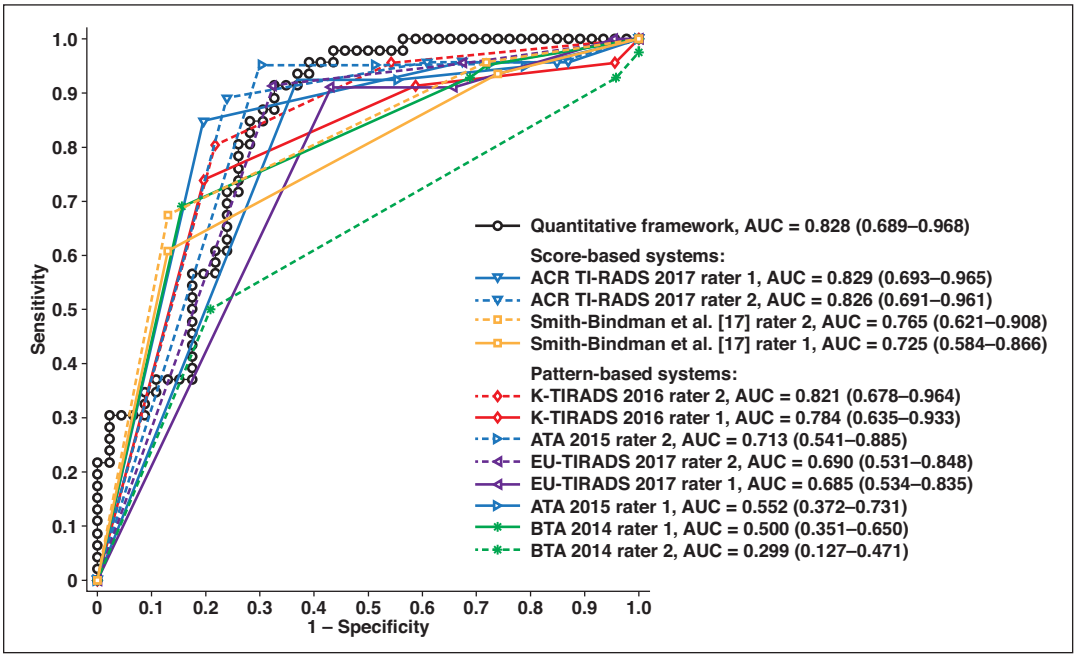

Fig. 2—Graph shows ROC curves for performance of proposed quantitative framework and implemented expert scoring systems. Values in parentheses are 95% CIs. ACR = American College of Radiology, TI-RADS and TIRADS = Thyroid Imaging Reporting and Data System, K = Korean, ATA = American Thyroid Association, EU = European Thyroid Association, BTA = British Thyroid Association.

TABLE 4: Performance of Classification Systems at Biopsy-Recommended Level in Terms of Point and Interval Estimates of Sensitivity and Specificity (*n* = 46)

Classifier	Sensitivity	Specificity	No. of Biopsies Spared	No. of Malignancies Missed
Quantitative, high sensitivity	1 (0.92–1)	0.43 (0.29–0.59)	20/46	0/46
Quantitative, high specificity	0.22 (0.11–0.36)	1 (0.92–1)	46/46	36/46
ACR TI-RADS 2017, rater 1	0.89 (0.76–0.96)	0.52 (0.37–0.67)	24/46	5/46
ACR TI-RADS 2017, rater 2	0.96 (0.85–0.99)	0.57 (0.41–0.71)	26/46	2/46
K-TIRADS 2016, rater 1	0.96 (0.85–0.99)	0.07 (0.01–0.18)	3/46	2/46
K-TIRADS 2016, rater 2	1 (0.92–1)	0.07 (0.01–0.18)	3/46	0/46
EU-TIRADS 2017, rater 1, high sensitivity	0.87 (0.74–0.95)	0.28 (0.16–0.43)	13/46	6/46
EU-TIRADS 2017, rater 1, high specificity	0.85 (0.71–0.94)	0.33 (0.20–0.48)	15/46	7/46
EU-TIRADS 2017, rater 2, high sensitivity	0.87 (0.74–0.95)	0.30 (0.18–0.46)	14/46	6/46
EU-TIRADS 2017, rater 2, high specificity	0.87 (0.74–0.95)	0.43 (0.29–0.59)	20/46	6/46
Smith-Bindman et al. [17], rater 1	0.15 (0.06–0.29)	0.98 (0.88–1)	45/46	39/46
Smith-Bindman et al. [17], rater 2	0.17 (0.08–0.31)	0.98 (0.88–1)	45/46	38/46
ATA 2015, rater 1, high sensitivity	1 (0.92–1)	0.11 (0.04–0.24)	5/46	0/46
ATA 2015, rater 1, high specificity	0.87 (0.74–0.95)	0.28 (0.16–0.43)	13/46	6/46
ATA 2015, rater 2, high sensitivity	1 (0.92–1)	0.17 (0.08–0.31)	8/46	0/46
ATA 2015, rater 2, high specificity	0.93 (0.82–0.99)	0.24 (0.13–0.39)	11/46	3/46
BTA 2014, rater 1, high sensitivity	0.96 (0.85–0.99)	0.20 (0.09–0.34)	9/46	2/46
BTA 2014, rater 1, high specificity	0.87 (0.74–0.95)	0.50 (0.35–0.65)	23/46	6/46
BTA 2014, rater 2, high sensitivity	0.98 (0.88–1)	0.00 (0.00–0.08)	0/46	1/46
BTA 2014, rater 2, high specificity	0.89 (0.76–0.96)	0.48 (0.33–0.63)	22/46	5/46

Note—For the proposed quantitative framework, two optimal points for high sensitivity and high specificity were calculated. For the pattern-based classifiers that had missing patterns, two decisions correspond to either biopsy of all unlabeled cases (high sensitivity) or no biopsy (high specificity). The ratios of spared biopsies of benign lesions and missed malignancies are given over the number of corresponding malignant or benign nodules considered by the stratification systems. Values in parentheses are 95% CIs. ACR = American College of Radiology, TI-RADS and TIRADS = Thyroid Imaging Reporting and Data System, K = Korean, EU = European Thyroid Association, ATA = American Thyroid Association, BTA = British Thyroid Association.

CI, 0.715–0.942). The analysis of difference in these values revealed insignificant differences between the proposed framework and ACR TI-RADS ($p = 0.99$) and the two raters ($p = 0.96$) (DeLong test).

Biopsy Decision

We compared the performances of the classifiers with respect to the recommendation for nodule biopsy. The performance of the quantitative framework for providing biopsy recommendation was validated at two cutoff points that reflected conservative or high-specificity and curative or high-sensitivity approaches to nodule management. For the pattern-based classification systems that did not provide a pattern for some nodules (ATA, BTA, and EU-TIRADS), such nodules were considered either all subjected to biopsy (high sensitivity) or no biopsy (high specificity).

When all methods were considered at their high-sensitivity biopsy decision cutoffs (no cancers missed), the quantitative framework had the highest sensitivity (1; 95% CI, 0.92–1), which was also achieved by rater 2 using K-TIRADS and by both raters using ATA, which was a statistically significant improvement over ACR TI-RADS for one rater, EU-TIRADS, and the system of Smith-Bindman et al. [17] for both raters. At the same time, the framework specificity of 0.43 (95% CI, 0.29–0.59) was statistically significantly higher than that of these three perfect sensitivity results. These specificities were 0.07 (95% CI, 0.01–0.18) for rater 2 using K-TIRADS, 0.11 (95% CI, 0.04–0.24) for rater 1 using ATA, and 0.17 (95% CI, 0.08–0.31) for rater 2 using ATA.

At the high-specificity cutoffs (no benign nodules biopsied), the highest specificity (1; 95% CI, 0.92–1) was achieved with the quantitative framework at sensitivity 0.22 (95% CI, 0.11–0.36). The specificity of the quantitative framework was statistically significantly higher than that for all other classifiers, except the system of Smith-Bindman et al. [17], which, however, had lower sensitivities of 0.15 (95% CI, 0.06–0.29) and 0.17 (95% CI, 0.08–0.31) for the two raters.

The performance summary for all of the implemented classifiers is shown in Table 4. As Figure 3 shows, the biopsy cutoffs of the six implemented systems lie either on the ROC curve of the quantitative framework or below it, indicating that for a chosen sensitivity or specificity, the framework can provide a decision with higher specificity and equal sensitivity.

Discussion

We present a quantitative framework for computerized stratification of risk of malignancy of thyroid nodules seen on US images. The predictive performance of the framework (AUC, 0.828) was better than or comparable to that of six classification systems implemented by annotations of two expert raters (AUC, 0.299–0.829). This finding suggests that the framework can provide expert-lev-

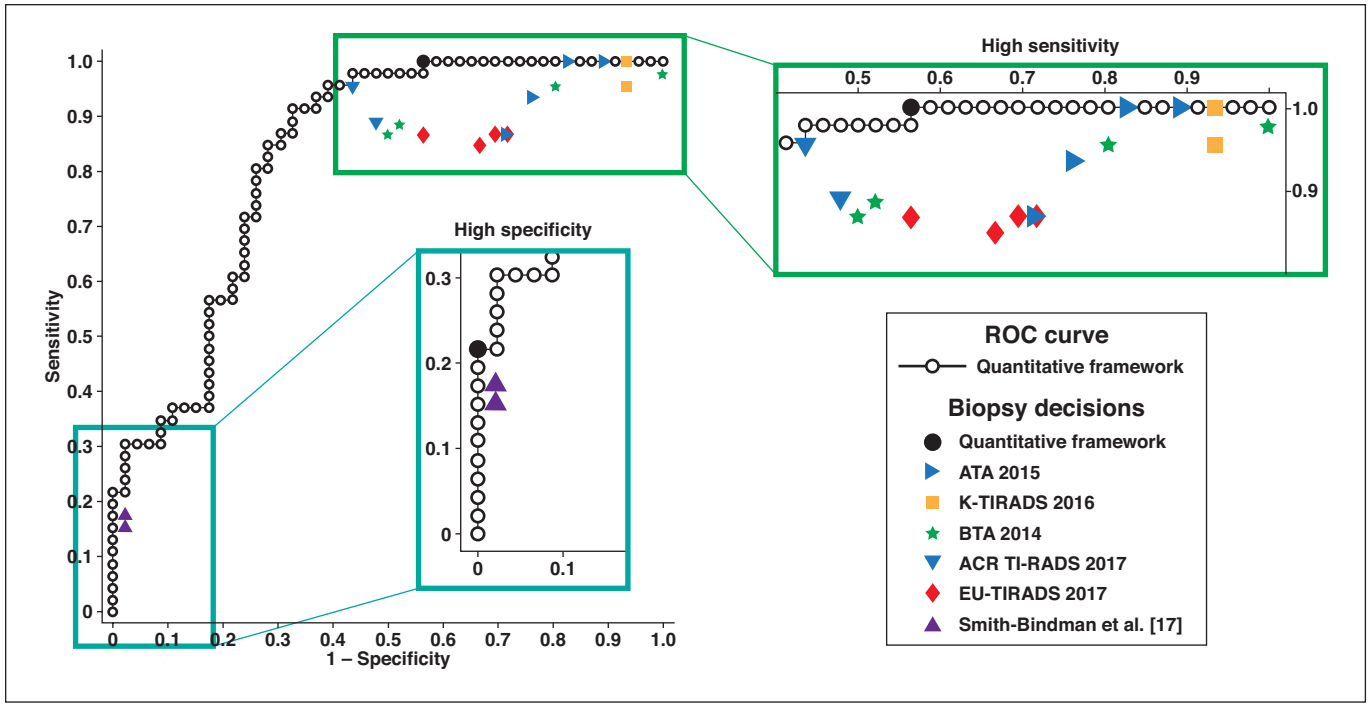

Fig. 3—Graph shows biopsy decision performance of six implemented systems for both raters and ROC curve of quantitative framework. For pattern-based classification and quantitative framework, two decision points for high sensitivity and high specificity approaches are plotted. ATA = American Thyroid Association, K = Korean, TI-RADS and TIRADS = Thyroid Imaging Reporting and Data System, BTA = British Thyroid Association, ACR = American College of Radiology, EU = European Thyroid Association.

el malignancy probabilities in an inexpensive and objective manner. Because different strategies for triage favor sensitivity or specificity and depending on the goal in a particular patient, one strategy may be favored over another. Further analysis of the potential biopsy recommendations revealed that the proposed framework could have spared 20 of 46 patients with benign nodules from biopsy in a curative strategy to triage (at sensitivity of 1) or identified 10 of 46 malignancies in a conservative strategy (at specificity of 1).

In the proposed framework, the quantitative features were selected to comprehensively encode the echogenicity, texture, edge sharpness, and margin curvature properties of thyroid nodules. This set of features also mimics the way a radiologist would annotate thyroid nodules when using systems such as ACR TI-RADS. However, some of these features rely on delineation of a nodule. When we used a subset of two features that are less dependent on the precision of manual segmentation—intensity and texture—we found a decrease in performance (AUC, 0.712). Therefore, we included the richer set of features that gave higher performance.

Although the choice of curative or conservative strategy of thyroid management remains subjective, there has been a shift toward the conservative approach because of the recognition that the incidence of thyroid cancer has tripled in the past 30 years without a concomitant change in mortality, suggesting a problem with cancer overdiagnosis and overtreatment [22, 23]. The more conservative strategy has been reflected in many of the more recent guidelines by the addition of constraints on the size of the nodule [6, 7, 9, 10]. In addition to comparing our quantitative framework with these classifiers, we also implemented the system described in the work of Smith-Bindman et al. [17], which was a prominent study proposing more stringent rules to reduce the number of unnecessary biopsies.

Solving the problem of thyroid cancer overdiagnosis has been the goal of many research groups and scientific committees. From the radiologist's perspective, a multitude of sonography-based scoring systems have been developed, most of them labeled TIRADS, to score malignancy risk and define a cutoff point for further diagnosis. The development of computer-aided tools to define the probability of malignancy is an active field of research in which methods are roughly divided into those that mimic radiologists' observations by defining sets of computational features and black-box methods whereby the diagnosis is learned directly, avoiding prediction of intermediate values as semantic features.

Several computer-aided methods have been proposed to predict or quantify sonographic semantic features [24–28] that could later be used as input to expert guidelines. For example, in their proposal for a computer-aided detection system, Chang. [26] inferred the presence of semantic features and followed the guidelines of the ATA [29] for decision making. Alternatively, some methods learn the classifiers on top of existing semantic features [30] or their combination with clinical variables [31]. For example, Wu et al. [30] found that their framework of radiologist-annotated semantic features and machine learning classifiers could provide results comparable to those of expert malignancy scoring.

Although a major objective in identification of benign nodules is to avoid unnecessary biopsies, some authors opt to predict the radiologist's annotation-based TIRADS malignancy score instead. For instance, Chi et al. [32], to determine the likelihood of nodule malignancy as defined in the TIRADS classification system of Kwak et al. [33], used a US image classification system based on a fine-tuned GoogLeNet model rather

than on pathologic proof. Although those authors reported accuracy of 0.99, the prediction of actual biopsy-proven malignancy is only as accurate as its TIRADS reference labels. In our dataset, for instance, the biopsy recommendation accuracy of the system of Kwak et al. was 0.55, but the highest accuracy among the six considered systems was 0.76. As such, even accurate prediction of a TIRADS malignancy risk category may not be ineffective in predicting the ultimate pathologic diagnosis.

The main limitation in comparing reported results among systems is that the evaluations were performed on different datasets of varying discrimination difficulty. For instance, using a dataset of 59 nodules, Chang et al. [34] reported an ROC AUC of 99% for their machine learning approach and ROC AUC of 98% for visual inspection by radiologists. However, the latter number may be indicative of a nonrepresentative dataset, because it is known that owing to overlap in sonographic features of benign and malignant nodules, the proportion of unnecessary biopsies of benign nodules is estimated to be as high as 93–95% [5].

To avoid misinterpretation of the performance statistics in a particular dataset, we considered it imperative to also provide results with expert systems in our dataset. When comparing the performance of the implemented systems with the results reported for other datasets, we as expected observed differences in absolute numbers that require adjusting the expectations for performance of a good computerized tool in a particular dataset. For instance, in our dataset, the performance of ACR TI-RADS was higher that that reported by Middleton et al. [35]. By comparing the performance of our framework with the performance of expert systems on our dataset rather than the reported values, we aimed to avoid the bias in our conclusions about the effectiveness of the framework. In addition, many of the previous studies were limited to training and testing in datasets as small as 20 patients [36]. On the other hand, expert-level performance has been found in deep learning methods in which large image datasets are used to train the models [37, 38]. Such approaches, however, can be prohibitive at institutions with low thyroid cancer prevalence, where it would take decades to assemble datasets with representative image samples for malignant nodule classes that such methods require. In that sense, our quantitative approach can be quickly and effectively adapted to the data of a particular clinic.

Thyroid cancer diagnosis currently relies on cytopathologic analysis of fine-needle biopsy specimens, which is invasive and anxiety-provoking for patients and may be nondiagnostic in as many as one-third of cases [17]. Identifying patients at low risk of cancer is important to avoid such unnecessary health risks and decrease societal health care costs. Attempts to address the issue by introducing more reliable prebiopsy cancer risk estimation with standard-of-care US images are based on radiologist-annotated standard descriptors, which are laborious to collect and are subject to interrater and interrater variability [31]. However, by extracting a rich set of features from these images in a computerized manner, our proposed framework expands the scope of features for analysis beyond merely those that are visible to the human eye and thereby maximizes use of the information available in the images. The use of computer-aided diagnostic tools based on frameworks similar to ours could improve the management of thyroid nodules and decrease the number of unnecessary biopsies and surgical risk while care is more appropriately directed at patients who need more invasive management.

Limitations

A limitation of this study was the use of only two principal projections, transverse and longitudinal, for extraction and analysis of semantic and computational features. Although the selection of these projections is standard in practice to evaluate nodule properties, such as shape and size, the use of multiple image frames from the US examinations could improve evaluation of other US features and will be analyzed in future work. Furthermore, the use of 3D transducers may provide more insight to the quantitative features of thyroid nodules, and the proposed framework can be generalized to such images in the future.

At present our method also has the limitation of requiring manual segmentation of nodules, which can be tedious and time-consuming. For this reason, we also explored the feasibility of eliminating the precise segmentation step by looking only at quantitative features that were independent of precise margin delineation: intensity and texture. In so doing, however, we found that the predictive performance decreased quite a bit. Thus, development of a segmentation-independent quantitative framework has been identified as a target for future investigations.

Through the use of a retrospective collection of US images that were routinely acquired at our clinic, our dataset will have certain heterogeneity factors. Although we excluded patients imaged with lower-frequency transducers (< 12 MHz) to minimize the acquisition variability and increase the image quality uniformity, the variability stemming from the multiple sonographers acquiring the images was unavoidable because more than 1000 studies over a 5-year period were evaluated. The impact of this variability is unlikely to be important because our US technologists are well-trained and American Registry for Diagnostic Medical Sonography certified, and we limited the analysis to the latest US equipment. For similar reasons, we did not exclude patients with background heterogeneity of the thyroid gland, such as those with Hashimoto thyroiditis or multinodular goiter. In all cases, we carefully selected the images of the nodule to correspond specifically to the nodule that was biopsied.

Another limitation was that our dataset of 92 nodules was selected with equal numbers of malignant and benign nodules, which is higher than the estimated prevalence of thyroid cancer and not truly representative of the case mix in our clinic. Although this would affect measures such as positive and negative predictive value, which were not used in our analyses, the reported AUC, sensitivity, and specificity values were not influenced. The small size of the dataset might have affected the absolute values of the methods considered; however, the main goals of our analysis were to study relative performance and to validate the proposed framework using established management guidelines that are currently used by a variety of international societies.

Conclusion

A quantitative framework for automated triage of thyroid nodules by use of sonography has been developed. The framework computes a rich set of computational features in the US images and estimates the probability of malignancy by use of an elastic net classifier. The performance of the framework in terms of AUC was 0.829, which was higher than or at the level of six expert radiologist annotation-based classifiers. For both curative and conservative treatment strategies, use of the proposed framework had the highest performance in terms of sensitivity and comparable performance in terms of specificity.

Given that our framework does not require qualitative assessments from an expert radiologist, it could reduce variation in practice.

Our results confirm the ultimate feasibility of computer-aided diagnostic systems for thyroid cancer risk estimation. Such systems could provide second-opinion malignancy risk estimation to clinicians and ultimately help decrease the number of unnecessary biopsies and surgical procedures. At present, however, a segmentation step to delineate the nodule boundary is critical to achieving expert performance, and manual segmentation is time-consuming. Future work will be directed at creating algorithms to accurately delineate the borders of nodules and developing a segmentation-independent quantitative framework. Once the algorithms and framework are devised, a fully automated process of thyroid nodule triage would be within reach.

References

1. Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *N Engl J Med* 2016; 375:614–617
2. Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Ann Intern Med* 1997; 126:226–231
3. Frates MC, Benson CB, Charboneau JW, et al.; Society of Radiologists in Ultrasound. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. *Radiology* 2005; 237:794–800
4. Howlader N, Noone A, Krapcho M, et al. SEER Cancer statistics review, 1975–2013, National Cancer Institute website. seer.cancer.gov/csr/1975_2013/. April 2016. Updated September 12, 2016. Accessed November 22, 2019
5. Jemal A, Murray T, Ward E, et al. Cancer statistics, 2005. *CA Cancer J Clin* 2005; 55:10–30
6. 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
7. 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
8. Perros P, Boelaert K, Colley S, et al.; British Thyroid Association. Guidelines for the management of thyroid cancer. *Clin Endocrinol (Oxf)* 2014; 81(suppl 1):1–122
9. Shin JH, Baek JH, Chung J, et al.; Korean Society of Thyroid Radiology (KSThR) and Korean Society of Radiology. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology consensus statement and recommendations. *Korean J Radiol* 2016; 17:370–395
10. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *Eur Thyroid J* 2017; 6:225–237
11. 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* 2018; 211:162–167
12. Grant EG, Tessler FN, Hoang JK, et al. Thyroid ultrasound reporting lexicon: white paper of the ACR Thyroid Imaging, Reporting and Data System (TIRADS) Committee. *J Am Coll Radiol* 2015; 12(12 part A):1272–1279
13. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated tracking of quantitative assessments of tumor burden in clinical trials. *Transl Oncol* 2014; 7:23–35
14. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern* 1973; 3:610–621
15. Tang X. Texture information in run-length matrices. *IEEE Trans Image Process* 1998; 7:1602–1609
16. Manay S, Cremers D, Hong BW, Yezzi AJ Jr, Soatto S. Integral invariants for shape matching. *IEEE Trans Pattern Anal Mach Intell* 2006; 28:1602–1618
17. 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
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005; 67:301–320
19. Sammut C, Webb GI, eds. Leave-one-out cross-validation. In: *Encyclopedia of machine learning*. Boston, MA: Springer, 2010:600–601
20. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; 33:1–22
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44:837–845
22. Hahn LD, Kunder CA, Chen MM, Orloff LA, Desser TS. Indolent thyroid cancer: knowns and unknowns. *Cancers Head Neck* 2017; 2:1
23. Brown RL, de Souza JA, Cohen EE. Thyroid cancer: burden of illness and management of disease. *J Cancer* 2011; 2:193–199
24. Chen KY, Chen CN, Wu MH, et al. Computerized detection and quantification of microcalcifications in thyroid nodules. *Ultrasound Med Biol* 2011; 37:870–878
25. Chen KY, Chen CN, Wu MH, et al. Computerized quantification of ultrasonic heterogeneity in thyroid nodules. *Ultrasound Med Biol* 2014; 40:2581–2589
26. Chang TC. The role of computer-aided detection and diagnosis system in the differential diagnosis of thyroid lesions in ultrasonography. *J Med Ultrasound* 2015; 23:177–184
27. Wu MH, Chen CN, Chen KY, et al. Quantitative analysis of echogenicity for patients with thyroid nodules. *Sci Rep* 2016; 6:35632
28. Choi YJ, Baek JH, Park HS, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. *Thyroid* 2017; 27:546–552
29. Cooper DS, Doherty GM, Haugen BR, et al.; American Thyroid Association (ATA) Guidelines Taskforce on Thyroid Nodules and Differentiated Thyroid Cancer. Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. *Thyroid* 2009; 19:1167–1214
30. Wu H, Deng Z, Zhang B, Liu Q, Chen J. Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography. *AJR* 2016; 207:859–864
31. Liu YI, Kamaya A, Desser TS, Rubin DL. A bayesian network for differentiating benign from malignant thyroid nodules using sonographic and demographic features. *AJR* 2011; 196:[web]W598–W605
32. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *J Digit Imaging* 2017; 30:477–486
33. Kwak JY, Han KH, Yoon JH, 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
34. Chang Y, Paul AK, Kim N, et al. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. *Med Phys* 2016; 43:554–567
35. 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* 2018; 210:1148–1154
36. Acharya UR, Vinitha Sree S, Krishnan MMR, Molinari F, Garberoglio R, Suri JS. Non-invasive automated 3D thyroid lesion classification in ultrasound: a class of ThyroScan™ systems. *Ultrasonics* 2012; 52:508–520
37. Wang L, Yang S, Yang S, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World J Surg Oncol* 2019; 17:12
38. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2019; 20:193–201