Current Problems in Diagnostic Radiology ▮ (2017) ▮▮▮–▮▮▮

Current Problems in Diagnostic Radiology

journal homepage: www.cpdrjournal.com

Diagnostic Performance of SRU and ATA Thyroid Nodule Classification Algorithms as Tested With a 1 Million Virtual Thyroid Nodule Model

Mitchell Boehnke, MD*, Nayana Patel, MD, Kristin McKinney, MD, Toshimasa Clark, MD

Division of Abdominal Imaging, Department of Diagnostic Radiology, The University of Colorado Anschutz Medical Campus, Aurora, CO

A B S T R A C T

Purpose: The Society of Radiologists in Ultrasound (SRU 2005) and American Thyroid Association (ATA 2009 and ATA 2015) have published algorithms regarding thyroid nodule management. Kwak et al. and other groups have described models that estimate thyroid nodules' malignancy risk. The aim of our study is to use Kwak's model to evaluate the tradeoffs of both sensitivity and specificity of SRU 2005, ATA 2009 and ATA 2015 management algorithms.

Materials and Methods: 1,000,000 thyroid nodules were modeled in MATLAB. Ultrasound characteristics were modeled after published data. Malignancy risk was estimated per Kwak's model and assigned as a binary variable. All nodules were then assessed using the published management algorithms. With the malignancy variable as condition positivity and algorithms' recommendation for FNA as test positivity, diagnostic performance was calculated.

Results: Modeled nodule characteristics mimic those of Kwak et al. 12.8% nodules were assigned as malignant (malignancy risk range of 2.0-98%). FNA was recommended for 41% of nodules by SRU 2005, 66% by ATA 2009, and 82% by ATA 2015. Sensitivity and specificity is significantly different (< 0.0001): 49% and 60% for SRU; 81% and 36% for ATA 2009; and 95% and 20% for ATA 2015.

Conclusion: SRU 2005, ATA 2009 and ATA 2015 algorithms are used routinely in clinical practice to determine whether thyroid nodule biopsy is indicated. We demonstrate significant differences in these algorithms' diagnostic performance, which result in a compromise between sensitivity and specificity.

Introduction

Thyroid cancer has led to 0.3% of all cancer deaths in 2016.[1] Fine-needle aspiration (FNA) is generally regarded as a safe and accurate diagnostic test to determine if a nodule is benign or malignant, but current management criteria lead to a large number of benign aspirations.[2,3] Sampling every thyroid nodule identified, regardless of size or sonographic features, would be impractical and costly as the prevalence of thyroid nodules is up to 65% in autopsy studies and the vast majority of lesions are not malignant.[4,5] Indeed, it is estimated that only 1 in 20 clinically identified nodules are in fact malignant, and only a subset of these malignancies are clinically significant.[6]

The Society of Radiology in Ultrasound in 2005 (SRU 2005) and the American Thyroid Association in 2009 and 2015 (ATA 2009 and ATA 2015) developed consensus recommendations that incorporate some of these clinical and radiological features to provide guidance on whether FNA should be performed.[7-9] Clinical features include vocal cord paralysis, regional lymphadenopathy, radiation exposure, family history, and male sex, all of which have been found to increase the risk of malignancy.[6] Imaging features include taller-than-wide morphology, microcalcifications, irregular margins, solidity, and hypoechogenicity.[5,10] Although larger nodule size has been associated with malignancy in some studies, no correlation was shown in others.[5,11] Doubling time similarly appears to be a poor indicator of malignancy.[12]

Multiple groups have developed risk models that predict the risk of malignancy based upon a given thyroid nodule's ultrasound characteristics.[5,13,14] We used the risk model as published by Kwak et al to test the diagnostic performance of SRU 2005, ATA 2009, and ATA 2015 management algorithms. Instead of drawing upon a data set of nodules clinically chosen for FNA, we used a virtual model that allows us to determine whether all of our modeled nodules are malignant or not. This in turn allowed us to calculate sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each algorithm. The aim of our study is to use Kwak's model to evaluate the tradeoffs of both sensitivity and specificity of SRU 2005, ATA 2009, and ATA 2015 management algorithms.

Materials and Methods

A total of 1,000,000 thyroid nodules were randomized based on differing characteristics modeled in MATLAB R2015b (MathWorks,

☆Dr Nayana Patel is scientific advisory board member for 3D Biopsy, LLC, but this entity was not involved in the research and there is no direct conflict of interest.

* Reprint requests: Mitchell Boehnke, MD, 12631 E 17th Avenue, Mail Stop 8200 Department of Radiology, University of Colorado School of Medicine Aurora, CO 80045.

E-mail address: Mitchell.boehnke@ucdenver.edu (M. Boehnke).

http://dx.doi.org/10.1067/j.cpradiol.2017.04.005
0363-0188/

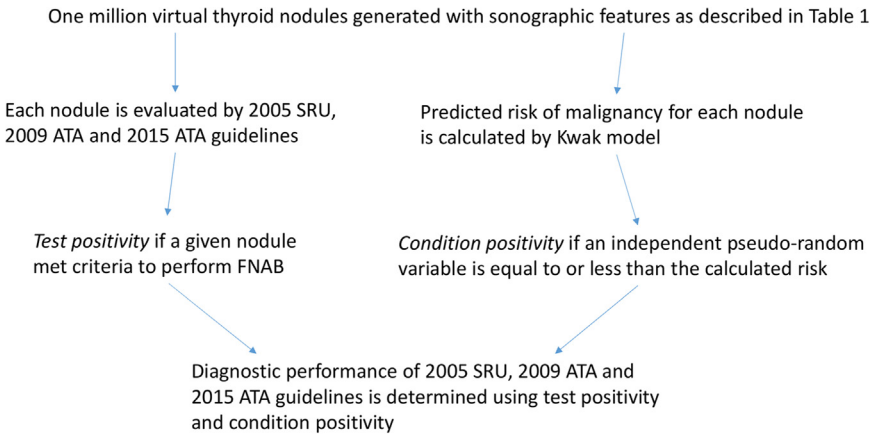

Fig. 1. Analysis flowchart. (Color version of figure is available online.)

Inc, Natick, MA). Size, solidity, echogenicity, margins, calcifications, and presence of taller-than-wide morphology were modeled after published data using independent pseudorandom variables.[5,15–17] To be more specific, a 1,000,000 × 6 matrix of independent pseudorandom variables was created using MATLAB's *rand* function, with each row corresponding to each modeled nodule, and 5 of the 6 the columns corresponding to modeled features. As an illustrative example, the literature indicates that 63.9% of nodules are solid, so nodule N would be designated as solid if the pseudorandom variable ($\in$ [0,1]) in the column determining solidity at row N was equal or less than 0.639.[5] Size was modeled in a slightly different manner, using the *exprndBounded* MATLAB function with coefficient 6.6 on the output range of 10-100 mm, as this was empirically demonstrated to create a size distribution that mimics that which has been previously described. We only modeled nodules 10 mm and greater in size because neither SRU 2005 nor ATA 2015 recommends FNA for any nodule below that threshold, rendering comparison irrelevant. ATA 2009 allowed for FNA for nodules as small as 5 mm if a patient had a high-risk history, but we did not account for a high-risk history in our model. The sixth column of pseudorandom variables was used in malignancy risk calculation. Malignancy risk was estimated for each nodule per Kwak model and assigned as a binary variable: calculated risk of N ($N \in [0,1]$) indicates malignancy if $M \leq N$ (pseudorandom variable $M \in [0,1]$). Independently, all nodules were assessed using SRU 2005, ATA 2009, and ATA 2015 algorithms. This analysis represented as a flowchart in Figure 1. With the binary malignancy variable as condition positivity and recommendation for FNA from each algorithm as test positivity, sensitivity, specificity, PPV, and NPV were calculated for each algorithm.

Table 1
Characteristics of modeled thyroid nodules

Number of nodules	1,000,000
Mean nodule diameter, mm	16.6
Median nodule diameter, mm	14.6
Range of nodule diameters, mm	10.0-98.1
% of nodules with solid composition	63.9%
% of nodules with mixed composition	36.1%
% of nodules with hypoechogenicity	40.7%
% of nodules with marked hypoechogenicity	4.9%
% of nodules with isoechogenicity or hyperechogenicity	54.4%
% of nodules with microlobulated margins	6.4%
% of nodules with irregular margins	9.7%
% of nodules with regular margins	83.9%
% of nodules with microcalcifications	9.7%
% of nodules with macrocalcifications	14.8%
% of nodules with no calcifications	75.5%
% of nodules with taller-than-wide morphology	11.9%

Chi-square analysis was performed using the Analysis Toolpak in Excel 2013 (Microsoft, Inc, Redmond, WA) and data were visualized using Excel and MATLAB. Due to multiple pairwise comparisons between the 3 groups, a threshold of 0.05/3 was used to determine statistical significance per the Bonferroni correction.

To allow for visualization of how typical combinations of imaging features (eg, hypoechoic but without microcalcification) would result in different risks of malignancies, we plotted a subset of imaging features against calculated malignancy risk via the model of Kwak et al in Excel, calculating the risk for each instance using MATLAB.

Results

A total of 1,000,000 thyroid nodules were modeled, with mean size 16.6 mm and median of 14.6 mm. Nodule characteristics mimic those described by Kwak et al. Risk of nodules' malignancy ranged from 2.0%-98% (mean 12.8% and median 6.8%). In total, 12.8% nodules were assigned as malignant. Thyroid nodule characteristics and calculated malignancy risk are listed in Tables 1 and 2, respectively.

FNA was recommended for 41% of nodules by SRU 2005, 66% of nodules by ATA 2009, and 82% by ATA 2015. Sensitivity and specificity of the algorithms was significantly different ($\chi^2 < 0.0001$ when compared to predicted values based off of test and condition positivity as defined in Figure 1): 49.3% and 60.0% for SRU; 81.4% and 36.4% for ATA 2009; and 95.4% and 20.5% for ATA 2015. Table 3 shows true/false positivity/negativity, PPV/NPV, sensitivity, and specificity for each algorithm.

Figure 2 illustrates calculated malignancy risk for 29 nodules that represent a subset of possible combinations of imaging features.

Table 2
Calculated malignancy risk and rate of test positivity for SRU, ATA 2009, and ATA 2015 thyroid nodule management algorithms

Mean calculated malignancy risk	12.8%
Median calculated malignancy risk	6.8%
Minimum calculated malignancy risk	2.0%
Maximum calculated malignancy risk	97.9%
% of nodules assigned as malignant via comparison with independent pseudorandom variable	12.8%
SRU % of nodules for which FNAB is recommended	41.1%
ATA 2009 % of nodules for which FNAB is recommended	65.9%
ATA 2015 % of nodules for which FNAB is recommended	81.6%

FNAB, fine-needle aspiration biopsy.