Surgery 165 (2019) 17–24

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

Interinstitutional variation in predictive value of the ThyroSeq v2 genomic classifier for cytologically indeterminate thyroid nodules[☆]

Andrea R. Marcadis, MD[a], Pablo Valderrabano, MD, PhD[b], Allen S. Ho, MD[c], Justin Tepe, BA, MPH[a], Christina E. Swartzwelder, RPA-C[a], Serena Byrd, MD[a], Wendy L. Sacks, MD[d], Brian R. Untch, MD[a], Ashok R. Shaha, MD[a], Bin Xu, MD, PhD[e], Oscar Lin, MD[e], Ronald A. Ghossein, MD[e], Richard J. Wong, MD[a], Jennifer L. Marti, MD[f,1,*], Luc G.T. Morris, MD, MSc[a,1,*]

[a] *Department of Surgery (Head and Neck Service), Memorial Sloan Kettering Cancer Center, New York, NY*
[b] *Department of Head and Neck Endocrine Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL*
[c] *Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, CA*
[d] *Department of Endocrinology, Cedars-Sinai Medical Center, Los Angeles, CA*
[e] *Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY*
[f] *Department of Surgery, New York Presbyterian/Weill Cornell Medical Center, New York, NY*

ARTICLE INFO

Article history:
Accepted 20 April 2018
Available online 22 October 2018

ABSTRACT

Background: The ThyroSeq v2 next-generation sequencing assay estimates the probability of malignancy in indeterminate thyroid nodules. Its diagnostic accuracy in different practice settings and patient populations is not well understood.

Methods: We analyzed 273 Bethesda III/IV indeterminate thyroid nodules evaluated with ThyroSeq at 4 institutions: 2 comprehensive cancer centers ($n = 98$ and 102), a multicenter health care system ($n = 60$), and an academic medical center ($n = 13$). The positive and negative predictive values of ThyroSeq and distribution of final pathologic diagnoses were analyzed and compared with values predicted by Bayes theorem.

Results: Across 4 institutions, the positive predictive value was 35% (22%–43%) and negative predictive value was 93% (88%–100%). Predictive values correlated closely with Bayes theorem estimates ($r^2 = 0.84$), although positive predictive values were lower than expected. *RAS* mutations were the most common molecular alteration. Among 84 *RAS*-mutated nodules, malignancy risk was variable (25%, range 10%–37%) and distribution of benign diagnoses differed across institutions (adenoma/hyperplasia 12%–85%, noninvasive follicular thyroid neoplasm with papillary-like nuclear features 5%–46%).

Conclusion: In a multi-institutional analysis, ThyroSeq positive predictive values were variable and lower than expected. This is attributable to differences in the prevalence of malignancy and variability in pathologist interpretations of noninvasive tumors. It is important that clinicians understand ThyroSeq performance in their practice setting when evaluating these results.

Crown Copyright © 2018 Published by Elsevier Inc. All rights reserved.

Introduction

Thyroid nodules with indeterminate cytologic characteristics on fine-needle aspiration (FNA) pose a management dilemma for clinicians, who aim to treat patients with malignant thyroid nodules while avoiding unnecessary medical and surgical therapy in patients with benign disease. Nodules classified as Bethesda III (atypia of undetermined significance/follicular lesion of undetermined significance) and Bethesda IV (follicular neoplasm/suspicious for follicular neoplasm) have estimated malignancy rates of 6% to 18% and 10% to 40%, respectively; however, these numbers vary markedly between institutions.[1-3] Because there is uncertainty regarding the probability of malignancy in indeterminate thyroid nodules (ITN), molecular tests are more often being used as tools to better triage patients to observation or surgery.

[☆] Accepted for oral presentation at the American Association of Endocrine Surgeons 39th Annual Meeting, Durham, NC, May 6-8, 2018.

[*] Corresponding author: Head and Neck Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, 1275 York Ave, C-1077, New York, NY 10065. Tel.: 212-639-3049, fax 212-717-3278.

E-mail address: morrisl@mskcc.org (L.G.T. Morris).

[1] Co-senior authors.

https://doi.org/10.1016/j.surg.2018.04.062
0039-6060/Crown Copyright © 2018 Published by Elsevier Inc. All rights reserved.

ThyroSeq v2 (CBLPath, Rye Brook, NY) is a commonly used molecular test to evaluate malignancy risk in ITN. This assay uses DNA and RNA extracted from FNA cytologic material to test for hotspot mutations in 14 genes (*AKT1*, *BRAF*, *CTNNB1*, *GNAS*, *HRAS*, *KRAS*, *NRAS*, *PIK3CA*, *PTEN*, *RET*, *TP53*, *TSHR*, *TERT* promoter, and *EIF1AX*) and 42 gene fusions involving *RET*, *PPARG*, *NTRK1*, *NTRK3*, *BRAF*, *IGF2BP3*, *ALK,* and *THADA*, as well as to detect expression changes in selected genes (including overexpression of *MET*, *PTH*, *TG*, and *TTF1*, among others).[4-6]

Several recent single-institution studies have been published that evaluate the performance of ThyroSeq v2.[4-8] These studies, however, are limited by their heterogeneous inclusion criteria and analysis methods as well as their single-institution nature, restricting our understanding of this assay's true diagnostic performance across different practice settings and patient populations. A diagnostic test is expected to have differing performance characteristics (positive and negative predictive values) and accuracy, depending on factors such as the prevalence of disease in the population.[9,10] Therefore the "real world" performance of ThyroSeq v2 as part of routine clinical care is likely to vary somewhat from institution to institution.

We have previously reported that a gene expression–based classifier for ITN exhibited widely variable performance across different institutions, likely attributable to differences in the prevalence of malignancy in different patient cohorts and variability in pathologist interpretation.[10] To better understand the performance of ThyroSeq v2 for ITN in routine clinical use across different practice settings and patient populations, we analyzed assay results and matched surgical pathologic specimens in patients from our institution and 3 other centers.

Methods

We performed a retrospective analysis of 273 Bethesda III or Bethesda IV ITN evaluated with ThyroSeq v2 and subsequently surgically resected, in 266 patients at 4 institutions. These included 98 ITN from a comprehensive cancer center (Memorial Sloan Kettering Cancer Center, New York, NY; MSKCC), 102 from a separate comprehensive cancer center (Moffitt Cancer Center, Tampa, FL; MCC), 13 from an academic medical center (Cedars-Sinai Medical Center, Los Angeles, CA; CSMC), and 60 from at a multihospital health care system (Mount Sinai Health System, New York, NY; MSHS). The ITN data from MSKCC and CSMC have not been previously published. The data from MCC and MSHS have been previously published, but both cohorts were reanalyzed for this study according to the inclusion criteria described later.[6,7] At MSKCC, CSMC, and MSHS, ThyroSeq v2 was ordered selectively for patients, or ordered by outside physicians before referral, whereas at MCC it was collected at the time of FNA and performed reflexively for all nodules with indeterminate cytology.[7] All specimens were collected between 2014 and 2017 (MSKCC 2014–2017, MCC 2014–2016, CSHS 2015–2017, MSHS 2014–2016). These years reflect the periods during which each institution conducted their internal reviews of ThyroSeq v2 performance.

All fine needle aspirations of the ITN were reviewed by fellowship-trained cytopathologists at the institution where the operation was performed (MSKCC, MCC, CSMC, or MSHS). Postoperatively, surgical pathology and preoperative ultrasound reports were re-reviewed and the biopsied nodule was correlated with findings on surgical pathologic examination by matching the nodule lobe, location within the lobe, and size. Incidental carcinomas independent from the biopsied nodule were considered separately. Seven of the patients included in the analysis had more than 1 ITN evaluated with ThyroSeq. The rates of indeterminate (Bethesda III and IV) category usage at each institution were MSKCC, 18% of all thyroid cytology specimens; MCC, 26%; CSMC, 15%; MSHS, 18%.[11-13]

ThyroSeq v2 results were considered ThyroSeq positive if alterations with malignancy probability >30% were reported. ITN with no genetic alterations identified, those exhibiting molecular alterations associated with <30% probability of malignancy, or those with low-frequency mutations corresponding to allelic fractions ≤5% (for *BRAF*, *TP53*, *AKT1*, *CTNB1*, *PIK3CA*, *TERT* promoter, and *RET*) or ≤10% (for *HRAS*, *KRAS*, *NRAS*, *PTEN*, *TSHR*, and *EIF1AX*) were considered to have a low risk for malignancy and were classified as ThyroSeq negative. The positive predictive values (PPVs) and negative predictive values (NPVs) of ThyroSeq results and distribution of final pathologic diagnoses were analyzed.

Given the reclassification of noninvasive, encapsulated follicular variant of papillary thyroid cancer as the nonmalignant entity noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) in 2017, all noninvasive, encapsulated, follicular variants of papillary thyroid carcinoma were re-reviewed by fellowship-trained head and neck pathologists and reclassified as NIFTP when appropriate.[14] In addition, because of the relatively recent reclassification of NIFTP as a nonmalignant entity, PPVs and NPVs were calculated in duplicate, with NIFTP alternatively considered benign or malignant. Measured PPVs and NPVs were compared with values predicted by Bayes theorem[15] based on published prevalence of malignancy of the indeterminate categories at each institution and the weighted test sensitivity and specificity of the Bethesda III and IV categories reported by Nikiforov et al.[4,5,11-13] To perform Bayesian analysis with NIFTP considered benign, the published malignancy prevalence for the indeterminate categories at each institution was adjusted by the proportion of nodules that were NIFTP at that institution. Statistical analysis was performed using GraphPad Version 7 (GraphPad Software, Inc, La Jolla, CA). This study was approved by the Institutional Review Board of Memorial Sloan Kettering Cancer Center.

Results

Of the 266 patients included, 75% were female (range, 71%–79%) with a mean age of 53 years (range, 42–56 years) (Table 1). Of 273 nodules, the mean size was 2.7 cm (2.4 cm Bethesda III, range 2.1–2.8 cm; 3.0 cm Bethesda IV, range 2.2–3.5 cm), and malignancy rates were 19% for Bethesda III nodules (range, 0%–35%) and 46% for Bethesda IV nodules (range, 20%–84%). Of 273 nodules, 155 (57%) had ThyroSeq positive results. The proportion of nodules with ThyroSeq-positive results was not significantly different between Bethesda III and Bethesda IV groups (59% vs 54%; $P = .37$). The ThyroSeq-negative group included 21 patients with low-frequency mutations or a quoted malignancy risk < 30% (7 MSKCC, 11 MCC, 0 CSMC, 3 MSHS) (Table 2).

The overall malignancy rate in ThyroSeq-positive nodules (PPV), with NIFTP considered nonmalignant ranged from 22% to 43% across institutions (Table 3). Overall the PPV of ThyroSeq v2 for malignancy was 35%. The NPV overall was 93% and ranged from 88% to 100% across institutions. The sensitivity was 87% (range, 73%–100%), and the specificity was 52% (20%–75%). Using the pretest probabilities for malignancy of the indeterminate categories at each institution, the predicted PPVs were 79% at MSKCC, 74% at MCC, 83% at CSMC, and 61% at MSHS. The PPVs were lower than predicted, whereas the NPVs were close to the expected numbers, and there was overall a strong correlation with predicted values ($r^2 = 0.84$) (Fig. 1). Because the sensitivity and specificity of ThyroSeq v2 initially reported by Nikiforov et al.[4,5] were determined before the reclassification of NIFTP as benign, the PPV and NPV of ThyroSeq v2 were recalculated with NIFTP considered malignant (Table 2, Fig. 1). Although this raised the PPVs, it also lowered the NPVs, leading to an overall weaker association with expected values ($r^2 = 0.73$).

Table 1
Demographic characteristics of patients included who underwent surgery for Bethesda III or IV indeterminate thyroid nodules with ThyroSeq v2 testing, by institution.

	MSKCC		MCC		CSMC		MSHS		Combined	
Patients (n)	97		97		13		59		266	
Female, n (%)	69 (71)		77 (79)		10 (77)		44 (75)		200 (75)	
Age in y, mean (SD)	51 (15)		56 (11)		42 (17)		52 (15)		53 (14)	
Nodules (n)	98		102		13		60		273	
	B III	B IV	B III	B IV	B III	B IV	B III	B IV	B III	B IV
Nodules, n (%)	55 (56)	43 (44)	52 (51)	50 (49)	9 (69)	4 (31)	45 (75)	15 (25)	161 (59)	112 (41)
Nodule size (cm), mean (SD)	2.1 (1.3)	2.9 (1.3)	2.6 (1.6)	3.0 (1.5)	2.8 (1.1)	2.2 (1.0)	2.5 (1.6)	3.5 (2.1)	2.4 (1.5)	3.0 (1.5)
ThyroSeq v2 positive, n (%)	41 (75)	33 (77)	16 (31)	17 (34)	8 (89)	3 (75)	30 (67)	7 (47)	95 (59)	60 (54)
Malignancy rate, n (%)	19 (35)	36 (84)	5 (10)	10 (20)	0	3 (75)	6 (13)	3 (20)	30 (19)	52 (46)

MSKCC, Memorial Sloan-Kettering Cancer Center; MCC, Moffitt Cancer Center; CSMC, Cedars-Sinai Medical Center; MSHS, Mount Sinai Health System; *B III*, Bethesda III; *B IV*, Bethesda IV; *SD*, standard deviation.

Table 2
Low allelic frequency and low malignancy risk mutations on ThyroSeq v2 considered to be ThyroSeq negative, by institution.

	MSKCC (n = *7*)	*MCC* (n = *11*)	*CSMC* (n = *0*)	*MSHS* (n = *3*)	***Malignant* n *(%)***
Benign mutations					
NIS overexpression (*n* = 1)	1				**0**
PTH expression (*n* = 2)		1		1	**0**
Low allelic frequency mutations					
EIF1AX (*n* = 2)	1	1			**0**
PTEN (*n* = 1)		1			**0**
RAS (*n* = 11)	4	6		1	**2 (18)**
TERT promoter (*n* = 1)	1				**0**
TP53 indeterminate[*] (*n* = 1)		1			**0**
TSHR (*n* = 2)		1		1	**0**
Malignant, *n* (%)	**1 (14)**	**1 (9)**	**0**	**0**	**2 (10)**

MSKCC, Memorial Sloan-Kettering Cancer Center; MCC, Moffitt Cancer Center; CSMC, Cedars-Sinai Medical Center; MSHS, Mount Sinai Health System.

[*] Indeterminate because of poor quality sequencing, suspicious for low allelic fraction mutation.

Table 3
Predictive values of ThyroSeq v2 for malignancy, by institution and combined, with NIFTP alternatively considered benign and malignant.

Institution	*ThyroSeq status*	*Malignant*	*NIFTP*	*Benign*	*Diagnostic performance (NIFTP benign)*	*Diagnostic performance (NIFTP malignant)*
MSKCC	ThyroSeq positive (*n* = 74)	32 (43%)	30 (41%)	12 (16%)	PPV 43%	PPV 84%
	ThyroSeq negative (*n* = 24)	3 (13%)	2 (8%)	19 (79%)	NPV 88%	NPV 79%
					Sensitivity 91%	Sensitivity 93%
					Specificity 33%	Specificity 61%
MCC	ThyroSeq positive (*n* = 33)	11 (33%)	5 (15%)	17 (52%)	PPV 33%	PPV 48%
	ThyroSeq negative (*n* = 69)	4 (6%)	5 (7%)	60 (87%)	NPV 94%	NPV 87%
					Sensitivity 73%	Sensitivity 64%
					Specificity 75%	Specificity 78%
CSMC	ThyroSeq positive (*n* = 11)	3 (27%)	1 (9%)	7 (64%)	PPV 27%	PPV 36%
	ThyroSeq negative (*n* = 2)	0 (0%)	0 (0%)	2 (100%)	NPV 100%	NPV 100%
					Sensitivity 100%	Sensitivity 100%
					Specificity 20%	Specificity 22%
MSHS	ThyroSeq positive (*n* = 37)	8 (22%)	2 (5%)	27 (73%)	PPV 22%	PPV 27%
	ThyroSeq-negative (*n* = 23)	1 (4%)	1 (4%)	21 (91%)	NPV 96%	NPV 91%
					Sensitivity 89%	Sensitivity 83%
					Specificity 43%	Specificity 44%
Overall	ThyroSeq positive (*n* = 155)	54 (35%)	38 (25%)	63 (41%)	PPV 35%	PPV 59%
	ThyroSeq negative (*n* = 118)	8 (7%)	8 (7%)	102 (86%)	NPV 93%	NPV 86%
					Sensitivity 87%	Sensitivity 85%
					Specificity 52%	Specificity 62%

Across all institutions, the most common molecular alterations encountered were mutations of the *RAS* genes, which were found in 62% of all nodules with positive ThyroSeq v2 results (Table 4). There were 96 nodules with *RAS* mutations (64 *NRAS*, 18 *HRAS*, 14 *KRAS*) of which 84 were nodules with isolated *RAS* mutations lacking any other molecular alteration (57 *NRAS*, 14 *HRAS*, 13 *KRAS*). The malignancy rate of all nodules with *RAS* mutations ranged from 9% to 41% (average 29%). Among *RAS*-mutant nodules, the rate of malignancy was significantly higher when additional molecular alterations were identified (*n* = 12) than when it was found in isolation (58% including only patients with *RAS* plus another mutation versus 25% with isolated *RAS* mutation; $P < .05$). Other common mutations and their malignancy rates included *BRAF V600E* (5/5; 100%), *BRAF K601E* (1/4; 25%), *EIF1AX* (isolated mutations 2/8; 25%; overall 7/19; 37%), *MET* overexpression (2/7; 29%), *PAX8/PPARG* fusion (5/11; 45%), and *THADA/IGF2BP3* fusion (3/8; 38%).

The considerable number of *RAS*-mutated nodules in this study allowed for close examination and comparison of their surgical

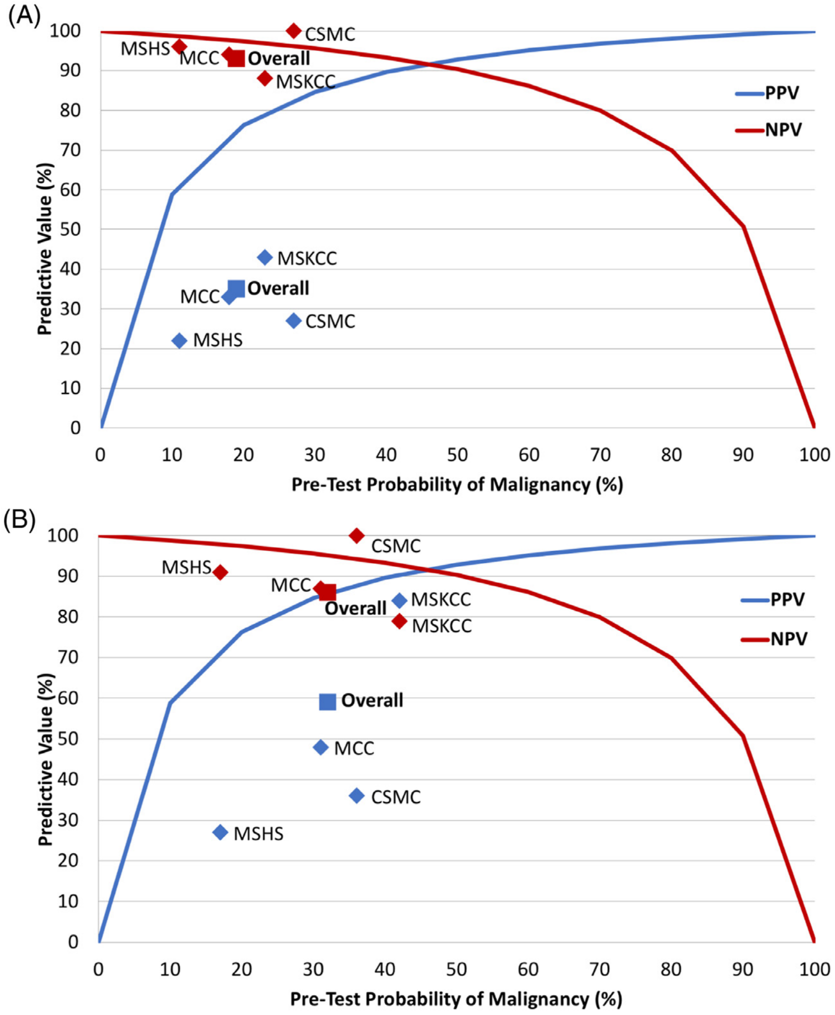


Fig. 1. (A) Positive and negative predictive values of ThyroSeq v2 compared with values predicted by Bayes theorem (pretest probability; MSKCC 23%, MCC 18%, CSMC 27%, MSHS 11%). (B) Alternative values with NIFTP considered malignant (pretest probability; MSKCC 42%, MCC 31%, CSMC 36%, MSHS 17%). The lines represent the values predicted by Bayes Theorem, and the points represent the actual positive and negative predictive values at each institution. *CSMC,* Cedars-Sinai Medical Center; *MCC,* Moffitt Cancer Center; *MSHS,* Mount Sinai Health System; *MSKCC,* Memorial Sloan-Kettering Cancer Center; *NIFTP,* noninvasive follicular thyroid neoplasm with papillary-like nuclear features; *NPV,* negative predictive value; *PPV,* positive predictive value.

pathologic diagnoses between institutions. At MSKCC, 37% of *RAS*-mutant nodules were malignant (17% classical variant papillary thyroid carcinoma [PTC], 10% follicular variant PTC, and 2.4% each of solid variant PTC, mixed classical and follicular variants PTC, follicular thyroid carcinoma, and Hurthle cell carcinoma). Other institutions had lower malignancy rates in *RAS*-mutated nodules: MCC 20%, CSMC 13%, and MSHS 10% (Fig. 2). Although the *RAS*-mutant nodules at MSKCC had a high rate of NIFTP diagnosis on final pathologic examination (46%), this rate was markedly lower at the other institutions (MCC 7%, CSMC 13%, MSHS 5%), where there were higher rates of other benign pathologic conditions, mostly follicular adenoma/nodular hyperplasia (MSKCC 12%, MCC 67%, CSMC 63%, MSHS 85%). Across all institutions, *RAS*-mutant nodules were most commonly follicular adenoma/nodular hyperplasia on surgical pathologic examination (44%), followed by NIFTP (26%) and classical variant PTC (11%).

Discussion

This multi-institutional analysis of ThyroSeq v2 diagnostic performance in 273 ITN reveals marked variation in test performance and distribution of pathologic diagnoses between institutions. Although overall test sensitivity was similar to what was originally reported (87% in our study vs 90% in original reports), the specificity was lower than reported (52% vs 93%).[4,5] This translated to a wide range of PPVs across institutions (22%–43%), all of which were substantially lower than the initially reported value of 81%[4,5] and generally lower than the reported probabilities of malignancy on the ThyroSeq report. The NPVs, on the other hand, were less variable (88%–100%), with the average NPV of 93% close to the reported value of 96%.[4,5] These findings of a high test NPV and sensitivity and relatively lower PPV and specificity reinforce the similar results found in one prior published series.[8]

In addition, we found wide variation in the prevalence of NIFTP diagnosis in *RAS*-mutant nodules (5%–46%) across institutions. These results are likely attributable to interobserver pathologist variability, and underscore the importance of studying the performance of molecular diagnostic assays employed as part of routine clinical care, across different institutions, and outside the more controlled settings afforded by investigational protocols at single institutions.

Even with the relatively low PPV of the test overall, there was a strong correlation with the predictive values estimated by Bayes theorem, with an r^2 value of 0.84. This indicates that the pretest probability of malignancy in ITN is an important determinant of PPV and NPV, suggesting that variable disease prevalence

Table 4
Risk of malignancy with ThyroSeq mutations (resected nodules) by institution.

	MSKCC	*MCC*	*CSMC*	*MSHS*	*Institutions combined*
ALK TD domain overexpression		0% (0/1)			0% (0/1)
BRAF V600E		100% (2/2)		100% (3/3)	100% (5/5)
BRAF K601E	0% (0/1)	50% (1/2)		0% (0/1)	25% (1/4)
BRAF L597V		100% (1/1)			100% (1/1)
BRAF deletion	0% (0/1)				0% (0/1)
BRAF K601E & *EIF1AX*	0% (0/1)				0% (0/1)
EIF1AX	67% (2/3)	0% (0/3)		0% (0/2)	25% (2/8)
EIF1AX + TSHR		0% (0/1)			0% (0/1)
ETV6/NTRK3				100% (1/1)	100% (1/1)
MET overexpression	0% (0/2)	50% (1/2)	100% (1/1)	0% (0/2)	29% (2/7)
NTRK3	100% (3/3)				100% (3/3)
PAX8/PPARG	50% (3/6)	100% (1/1)	0% (0/1)	33% (1/3)	45% (5/11)
HRAS Q61 (K, R)	13% (1/8)	50% (1/2)	0% (0/2)	0% (0/2)	14% (2/14)
KRAS Q61 (K, R), G12V, G12D	40% (2/5)	0% (0/3)	50% (1/2)	33% (1/3)	31% (4/13)
NRAS Q61 (K, R), G13R	43% (12/28)	20% (2/10)	0% (0/4)	7% (1/15)	26% (15/57)
Isolated *RAS* mutations	**37% (15/41)**	**20% (3/15)**	**13% (1/8)**	**10% (2/20)**	**25% (21/84)**
HRAS & calcitonin expression	100% (1/1)				100% (1/1)
HRAS & *EIF1AX*	100% (1/1)			0% (0/2)	33% (1/3)
KRAS & *EIF1AX*	0% (0/1)				0% (0/1)
NRAS & *EIF1AX*	67% (2/3)				67% (2/3)
NRAS & *TERT* promoter	0% (0/1)		100% (1/1)		50% (1/2)
NRAS & *TERT* promoter & *EIF1AX*	100% (1/1)	100% (1/1)			100% (2/2)
All *RAS* mutations	**41% (20/49)**	**25% (4/16)**	**22% (2/9)**	**9% (2/22)**	**29% (28/96)**
RET/PTC1	100% (1/1)			100% (1/1)	100% (2/2)
TERT promoter	0% (0/1)				0% (0/1)
THADA/IGF2BP3	50% (2/4)	33% (1/3)		0% (0/1)	38% (3/8)
TP53	100% (1/1)			0% (0/1)	50% (1/2)
TSHR	0% (0/1)	0% (0/1)			0% (0/2)
Mutations combined	**43% (32/74)**	**33% (11/33)**	**27% (3/11)**	**22% (8/37)**	**35% (54/155)**

NIFTP was considered benign to calculate the rates of malignancy.
MSKCC, Memorial Sloan-Kettering Cancer Center; MCC, Moffitt Cancer Center; CSMC, Cedars-Sinai Medical Center; MSHS, Mount Sinai Health System.

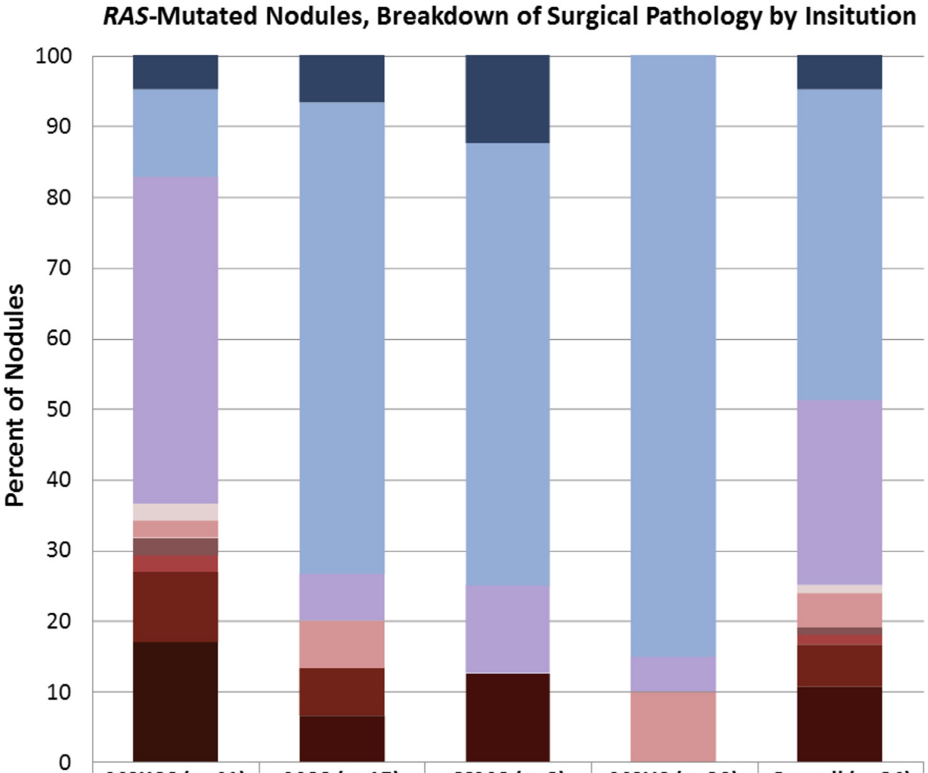
is a major cause of variable test performance. When the PPV and NPV were recalculated with NIFTP considered malignant, as it was when the test was designed, there was a weaker correlation with Bayes theorem predictions, with an r^2 of 0.73. This appears to be attributable to a lowering of the NPV, with a smaller increase in PPV, if NIFTP are considered malignant.

The most commonly detected mutations in this series were the *RAS* mutations, which had a malignancy rate of 25% as an isolated mutation and 29% when patients with all *RAS* mutations (including those with RAS plus additional mutations) were included in the analysis. These values are much lower than the estimated probabilities of malignancy in the ThyroSeq reports—most of these mutations have stated malignancy rates of >70%–80%.

There was wide variation in the prevalence of NIFTP in resected *RAS*-mutant ITN across institutions, ranging from 5% to 46%. At MSKCC, NIFTP was the most common overall and nonmalignant entity in *RAS*-mutated nodules. This was substantially higher than the NIFTP rate at the 3 other institutions, where the most common overall and benign diagnosis was follicular adenoma/nodular hyperplasia. Even among the other 3 institutions (MCC, CSMC, MSHS) there was significant variation in the NIFTP (5%–13%) and follicular adenoma/nodular hyperplasia (63%–85%) diagnosis rates. Although true differences in histologic diagnosis cannot be ruled out because of the absence of a centralized pathology review, these results suggest differences in histologic interpretation of encapsulated noninvasive follicular lesions as the most likely explanation, which is a well-recognized phenomenon.[14,16,17]

Even before the designation of NIFTP, it had been found that significant inter- and intra-observer variation existed in distinguishing follicular variant of papillary thyroid carcinoma (FVPTC) from follicular adenoma and follicular thyroid cancer, mostly relying on the identification of nuclear features of papillary carcinoma (pseudo-inclusions, nuclear grooves, ground-glass nuclei) within the lesion, which is not always straightforward. Studies have reported low rates of concordance (39%) among pathologists in the diagnosis of FVPTC.[17] Strikingly, the 24 expert thyroid pathologists convened to reclassify encapsulated FVPTC initially agreed on only 1 of 138 cases.[14] The addition of NIFTP as a distinct category of neoplasm with follicular and papillary features added an additional layer of complexity to this diagnosis. To attempt to mitigate this interpathologist variation, clear histopathologic criteria for NIFTP diagnosis have been established, including a 3-point “nuclear score” and other features, such as encapsulation or clear demarcation, follicular growth pattern with <1% papillae, no psammoma bodies, <30% solid, trabecular, or insular growth pattern, no vascular or capsular invasion, and no tumor necrosis or high mitotic activity.[14] Even with these strict criteria and with pathologist training, there remains marked interpathologist variation in the diagnosis of NIFTP versus follicular adenoma.[14] Interobserver variability in NIFTP diagnosis was the major source of interinstitutional variability, even among 4 institutions with high volumes of surgical cases of thyroid disease and fellowship-trained specialty pathologists. It is possible that this variation may be wider outside of institutions with high-volume thyroid surgical pathology.

Currently there is debate as to whether NIFTP is a premalignant or benign entity[14,18]. NIFTP may represent an in situ carcinoma or hyperplastic proliferation, with evidence of the natural history of untreated NIFTP currently lacking. Unlike other premalignant lesions, it is not clear that there are any well-documented cases of NIFTP developing metastatic disease, even at a low rate.[14] Interestingly, *RAS* mutations are found in follicular adenomas and nodular hyperplasia, similar to the high prevalence of *BRAF* mutations in benign nevi arising in the skin.[19] At present this issue is unresolved, and it remains unknown whether all NIFTPs need to be surgically treated. If we consider NIFTP benign and not requiring definitive diagnosis or resection, the percentage of *RAS*-mutated nodules ultimately requiring surgery would be 10% to 37% (the carcinomas). On the other hand, if we consider NIFTP pre-



	MSKCC (n=41)	MCC (n=15)	CSMC (n=8)	MSHS (n=20)	Overall (n=84)
■ Hurthle cell adenoma	4.9	6.7	12.5	0	4.8
■ Follicular adenoma / Nodular hyperplasia	12.2	66.7	62.5	85.0	44.0
■ NIFTP	46.3	6.7	12.5	5.0	26.2
■ Hurthle cell carcinoma	2.4	0	0	0	1.2
■ Follicular thyroid carcinoma	2.4	6.7	0	10.0	4.8
■ PTC solid variant	2.4	0	0	0	1.2
■ PTC classical and follicular variants	2.4	0	0	0	1.2
■ PTC follicular variant	9.8	6.7	0	0	6.0
■ PTC classical variant	17.1	6.7	12.5	0	10.7

Fig. 2. Breakdown of *RAS*-mutated nodules by surgical pathologic diagnosis and institution; numbers listed as percentages. *CSMC,* Cedars-Sinai Medical Center; *MCC,* Moffitt Cancer Center; *MSHS,* Mount Sinai Health System; *MSKCC,* Memorial Sloan-Kettering Cancer Center; *NIFTP,* noninvasive follicular thyroid neoplasm with papillary-like nuclear features; *PTC,* papillary thyroid carcinoma.

malignant and requiring surgical resection, the percentage of *RAS*-mutant cases requiring surgery would vary dramatically across institutions, ranging from 15% to 83%.

Some of the limitations of this study include its retrospective and multi-institutional design. Although efforts were made to exclude incidental microcarcinomas in the surgical pathologic specimens, the inadvertent inclusion of these would artificially inflate the PPV. In addition, intra- and inter-institutional differences exist in criteria for ordering ThyroSeq v2 and in interpretation of both cytologic and surgical pathologic information. In 3 institutions in this study (MSKCC, CSMC, and MSHS), ThyroSeq v2 was ordered selectively for nodules in which results would potentially change management. Selective use of a molecular assay may lower the reported PPV and NPV of the test because nodules that are more likely to be malignant and test positive on ThyroSeq v2 would be triaged directly to surgical resection (bypassing ThyroSeq testing), and those more likely to be benign and testing negative on ThyroSeq v2 would be observed (also bypassing ThyroSeq testing). Though the extent of this effect is unknown, it may contribute to some of the deviation of the predictive values in this study from the values expected on Bayesian analysis. This study sought to audit the performance of this assay in different patient populations and clinician and pathologist practices, and to therefore reflect “real world” use of this assay. The wide interinstitutional variability in performance is likely attributable to these factors.

This analysis of ThyroSeq v2 performance is the largest and most comprehensive since the test’s introduction in 2014, and it helps to reveal inherent differences in test performance between institutions. This variation in test performance is likely attributable to differences in disease prevalence, selection of nodules for molecular testing, and pathologist interpretation of resected nodules. These factors exemplify the importance of distinguishing *efficacy* (results of an intervention or diagnostic test under ideal circumstances) from *effectiveness* (results identified in “real world” clinical practice). It is not uncommon for the performance of a diagnostic test to be reduced in clinical practice compared with the initial results reported in highly controlled studies. As newer versions of ThyroSeq and other molecular tests are marketed and used more widely, it is important that physicians are proficient in understanding and interpreting these data in the context of the PPV and NPV in their own practice setting, in order to correctly interpret the results and provide optimal patient care.

References

1. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid*. 2017;27:1341–1346.
2. Ho AS, Sarti EE, Jain KS, Wang H, Nixon IJ, Shaha AR, et al. Malignancy rate in thyroid nodules classified as Bethesda category III (AUS/FLUS). *Thyroid*. 2014;24:832–839.
3. Straccia P, Rossi ED, Bizzarro T, Brunelli C, Cianfrini F, Damiani D, et al. A meta-analytic review of the Bethesda System for Reporting Thyroid Cytopathology: has the rate of malignancy in indeterminate lesions been underestimated? *Cancer Cytopathol*. 2015;123:713–722.
4. Nikiforov YE, Carty SE, Chiosea SI, Coyne C, Duvvuri U, Ferris RL, et al. Highly accurate diagnosis of cancer in thyroid nodules with follicular neoplasm/suspicious for a follicular neoplasm cytology by ThyroSeq v2 next-generation sequencing assay. *Cancer*. 2014;1(120):3627–3634.
5. Nikiforov YE, Carty SE, Chiosea SI, Coyne C, Duvvuri U, Ferris RL, et al. Impact of the multi-gene ThyroSeq next-generation sequencing assay on cancer diagnosis in thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance cytology. *Thyroid*. 2015;25:1217–1223.
6. Taye A, Gurciullo D, Miles BA, Gupta A, Owen RP, Inabnet WB, et al. Clinical performance of a next-generation sequencing assay (ThyroSeq v2) in the evaluation of indeterminate thyroid nodules. *Surgery*. 2018;163:97–103.
7. Valderrabano P, Khazai L, Leon ME, Thompson ZJ, Ma Z, Chung CH, et al. Evaluation of ThyroSeq v2 performance in thyroid nodules with indeterminate cytology. *Endocr Relat Cancer*. 2017;24:127–136.
8. Shrestha RT, Evasovich MR, Amin K, Radulescu A, Sanghvi TS, Nelson AC, et al. Correlation between histological diagnosis and mutational panel testing of thyroid nodules: a two-year institutional experience. *Thyroid*. 2016;26:1068–1076.
9. Ferris RL, Baloch Z, Bernet V, Chen A, Fahey TJ, Ganly I, et al. American Thyroid Association statement on surgical application of molecular profiling for thyroid nodules: current impact on perioperative decision making. *Thyroid*. 2015;25:760–768.
10. Marti JL, Avadhani V, Donatelli LA, Niyogi S, Wang B, Wong RJ, et al. Wide inter-institutional variation in performance of a molecular classifier for indeterminate thyroid nodules. *Ann Surg Oncol*. 2015;22:3996–4001.
11. Iskandar ME, Bonomo G, Avadhani V, Persky M, Lucido D, Wang B, et al. Evidence for overestimation of the prevalence of malignancy in indeterminate thyroid nodules classified as Bethesda category III. *Surgery*. 2015;157:510–517.
12. Sacks WL, Bose S, Zumsteg ZS, Wong R, Shiao SL, Braunstein GD, et al. Impact of Afirma gene expression classifier on cytopathology diagnosis and rate of thyroidectomy. *Cancer Cytopathol*. 2016;124:722–728.
13. Valderrabano P, Leon ME, Centeno BA, Otto KJ, Khazai L, McCaffrey JC, et al. Institutional prevalence of malignancy of indeterminate thyroid cytology is necessary but insufficient to accurately interpret molecular marker tests. *Eur J Endocrinol*. 2016;174:621–629.
14. Nikiforov YE, Seethala RR, Tallini G, Baloch ZW, Basolo F, Thompson LDR, et al. Nomenclature revision for encapsulated follicular variant of papillary thyroid carcinoma: a paradigm shift to reduce overtreatment of indolent tumors. *JAMA Oncol*. 2016;2:1023–1029.
15. Sox HC. Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. *Ann Intern Med*. 1986;104:60–66.
16. Lloyd RV, Erickson LA, Casey MB, Lam KY, Lohse CM, Asa SL, et al. Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. *Am J Surg Pathol*. 2004;28:1336–1340.
17. Elsheikh TM, Asa SL, Chan JKC, DeLellis RA, Heffess CS, LiVolsi VA, et al. Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. *Am J Clin Pathol*. 2008;130:736–744.
18. Kim TH, Lee M, Kwon A-Y, Choe J-H, Kim J-H, Kim JS, et al. Molecular genotyping of the non-invasive encapsulated follicular variant of papillary thyroid carcinoma. *Histopathology*. 2018;72:948–961.
19. Roh MR, Eliades P, Gupta S, Tsao H. Genetics of melanocytic nevi. *Pigment Cell Melanoma Res*. 2015;28:661–672.

Discussion

Dr. Sareh Parangi (Boston, MA): I just have a practical question. For the patient who comes into my office and has a *RAS* mutation, and if I am in the camp who doesn't consider NIFTP malignant, what's the bottom line that I can tell them? They might have this follicular cancer (something really aggressive) that they need a completion thyroidectomy for? What did you find? What percent of the patients had a real follicular thyroid cancer or something more aggressive, or some other thing that would cause you to do a total thyroidectomy?

Dr. Andrea R. Marcadis: I think the answer to your question is that it depends on the institution that you come from. If you look at the rates of true malignancies in *RAS* mutant nodules, we see that there is a lot of variability between institutions. At one institution, it was about 10%, in another it was almost 40%. So I think that finding really underscores the fact that you need to understand the performance of the test and your underlying rates of malignancy in your population in order to best interpret these results.

Dr. Sareh Parangi (Boston, MA): So even if we added up all the follicular thyroid cancers, 10%, and the follicular variant PTC, 10%, and the maximum of classic variant PTC, it would still be lower than 40%? And it's probably considerably lower because if these are classic PTC that are not too big, you wouldn't worry. Thank you.

Dr. Allan Siperstein (Cleveland, OH): The purpose of these tests is obviously an adjunct to help us to determine who to operate on and who not to operate on. We obviously have additional clinical information. I wonder if you have overlaid this with primitive things such as nodule size, and more sophisticated things like sonographic TI-RAD score.

Dr. Andrea R. Marcadis: Because this was a multi-institutional study, we didn't go too far into the specific ultrasound characteristics recorded at each institution because there was variability in the way that these ultrasound characteristics were interpreted. We did record nodule size for all patients and average size was similar between institutions, but this was not the focus of our study.

Dr. John Lew (Miami, FL): Dr. Marcadis, congratulations to you and your colleagues for a nice presentation of a timely study. You really illustrate the importance of looking to one's own institution regarding these testing modalities. As you show, there's a real variation in terms of results. I noticed in your slides that you saw differences in NIFTP rates among institutions as well as your positive predictive value. Any thoughts on why you are seeing such variation in that?

Dr. Andrea R. Marcadis: I think there could be a few different reasons why we are seeing these differences in NIFTP rates between institutions. One of the reasons that we can't exclude is that there may be true difference in this NIFTP entity between institutions. We didn't have a centralized review so we can't exclude that entirely.

The second and I think more likely reason is that there are differences in the way that the pathologists at these different institutions interpret the same histology. This is well documented in the literature particularly for follicular neoplasms—what one pathologist may call a follicular adenoma, another may call a NIFTP, and yet another a follicular variant of papillary thyroid cancer. And so I think this likely played a role in the differences we observed in NIFTP rate between institutions and is actually a subject of great interest and ongoing investigation at our institution.

Dr. Quan-Yang Duh (San Francisco, CA): A brief comment and a quick question. Many years ago, we actually showed that if you add *RAS* mutation to your panels, you would increase sensitivity but would decrease the specificity, which is what you are showing right now. And it is surprising that more than 60% of the mutations that you find are *RAS* mutations.

The question that I have for you is in contrast to Allan Siperstein, who wanted to spend less money. What if you were to spend more money and also do an Afirma test on these *RAS*-mutated lesions? What would you find?

Dr. Andrea R. Marcadis: I think, actually, that the Afirma test and the ThyroSeq test are similar in the way that they perform, in that they both have a high negative predictive value and relatively lower positive predictive value. I don't think it would add much. With the Afirma test, we don't see the particular gene that's mutated so I think that ThyroSeq may add value in that aspect, but otherwise they perform similarly.

Dr. Quan-Yang Duh (San Francisco, CA): So a patient that came to me with both tests done, a nurse, with a 0.8-centimeter lesion, and RAS positive, Afirma low risk. Would you operate on that patient?

Dr. Andrea R. Marcadis: As always, it depends on the patient and the institution, but no, probably not.

Dr. Avital Harari (Los Angeles, CA): I think it's similar to what we are seeing on the West Coast actually. My question is more about the negative predictive value. As you know, with Afirma, they had a pretty good surgical cohort and ThyroSeq did not have a great surgical cohort to confirm. So my question is for your study, how did you confirm your negative predictive value?

Dr. Andrea R. Marcadis: I know a lot of studies have looked at the nodules that had benign molecular results, and if those nodules weren't resected, they still considered them benign in the calculation of the negative predictive value. For our study, we only looked at surgically resected nodules. We made sure to match the nodule that was biopsied to where the lesion was on final pathology. Everything was surgically confirmed.

Dr. Linwah Yip (Pittsburgh, PA): Congratulations on an excellent study and it was very beautifully presented. I want to highlight that I think what your study design is novel and I think really underscores the importance of these clinical efficacy studies for these tests.

I have 2 questions for you.

The first is, I was wondering if you had any idea what the selection criteria was for obtaining molecular tests in these biopsies. As Dr. Duh pointed out, an 8-millimeter indeterminate nodule probably shouldn't have been biopsied anyway, but potentially maybe some of those selection issues contributed to the things that we are seeing here that would have been unusual in your results.

My second question is if you have any idea how many of those nodules that were positive or negative actually went to surgery. If there were a lot of negative nodules that were observed, the conclusions may not be 100% accurate.

Dr. Andrea R. Marcadis: Many of the patients were referred by their primary care physicians or endocrinologists with the ThyroSeq test already performed, and so for those patients we don't know what criteria the outside clinicians used when ordering the test. The selection criteria for obtaining molecular tests depended somewhat on the participating institution as well. For example at some institutions included in our study (particularly Moffitt), ThyroSeq was performed on all indeterminate nodules, while at the other institutions, it was more selective based on if it was thought that it would help to guide management. So there was a little bit of difference there, but I think these differences are reflective of the real world use of the ThyroSeq test.

And to answer your question about how many of the benign nodules actually went to surgery, for this study, we only included nodules that went to surgery, so all of them. We didn't follow the patients that had molecular tests that didn't go to surgery, but it's something that we can look at in future studies.