JAMA Surgery | Original Investigation

Performance of a Genomic Sequencing Classifier for the Preoperative Diagnosis of Cytologically Indeterminate Thyroid Nodules

Kepal N. Patel, MD; Trevor E. Angell, MD; Joshua Babiarz, PhD; Neil M. Barth, MD; Thomas Blevins, MD; Quan-Yang Duh, MD; Ronald A. Ghossein, MD; R. Mack Harrell, MD; Jing Huang, PhD; Giulia C. Kennedy, PhD; Su Yeon Kim, PhD; Richard T. Kloos, MD; Virginia A. LiVolsi, MD; Gregory W. Randolph, MD; Peter M. Sadow, MD, PhD; Michael H. Shanik, MD; Julie A. Sosa, MD; S. Thomas Traweek, MD; P. Sean Walsh, MPH; Duncan Whitney, PhD; Michael W. Yeh, MD; Paul W. Ladenson, MD

IMPORTANCE Use of next-generation sequencing of RNA and machine learning algorithms can classify the risk of malignancy in cytologically indeterminate thyroid nodules to limit unnecessary diagnostic surgery.

OBJECTIVE To measure the performance of a genomic sequencing classifier for cytologically indeterminate thyroid nodules.

DESIGN, SETTING, AND PARTICIPANTS A blinded validation study was conducted on a set of cytologically indeterminate thyroid nodules collected by fine-needle aspiration biopsy between June 2009 and December 2010 from 49 academic and community centers in the United States. All patients underwent surgery without genomic information and were assigned a histopathology diagnosis by an expert panel blinded to all genomic information. There were 210 potentially eligible thyroid biopsy samples with Bethesda III or IV indeterminate cytopathology that constituted a cohort previously used to validate the gene expression classifier. Of these, 191 samples (91.0%) had adequate residual RNA for validation of the genomic sequencing classifier. Algorithm development and independent validation occurred between August 2016 and May 2017.

EXPOSURES Thyroid nodule surgical histopathology diagnosis by an expert panel blinded to all genomic data.

MAIN OUTCOMES AND MEASURES The primary end point was measurement of genomic sequencing classifier sensitivity, specificity, and negative and positive predictive values in biopsies from Bethesda III and IV nodules. The secondary end point was measurement of classifier performance in biopsies from Bethesda II, V, and VI nodules.

RESULTS Of the 183 included patients, 142 (77.6%) were women, and the mean (range) age was 51.7 (22.0-85.0) years. The genomic sequencing classifier had a sensitivity of 91% (95% CI, 79-98) and a specificity of 68% (95% CI, 60-76). At 24% cancer prevalence, the negative predictive value was 96% (95% CI, 90-99) and the positive predictive value was 47% (95% CI, 36-58).

CONCLUSIONS AND RELEVANCE The genomic sequencing classifier demonstrates high sensitivity and accuracy for identifying benign nodules. Its 36% increase in specificity compared with the gene expression classifier potentially increases the number of patients with benign nodules who can safely avoid unnecessary diagnostic surgery.

Invited Commentary page 824

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Kepal N. Patel, MD, Division of Endocrine Surgery, Department of Surgery, New York University Langone Medical Center, 530 First Ave, Ste 6H, New York, NY 10016 (kepal.patel@nyumc.org).

JAMA Surg. 2018;153(9):817-824. doi:10.1001/jamasurg.2018.1153
Published online May 23, 2018.

Thyroid cancer incidence has increased substantially in the United States in recent decades, with evidence to support both an increase in detection[1] and a true increase in occurrence.[2] Thyroid nodules are palpable in 5% of adults[3] and are visualized with contemporary imaging in more than one-third of adults.[3-5] Malignancy is present in only 5% to 15% of all thyroid nodules,[3,5-7] and definitive diagnosis is achieved by surgical histopathology on resected tissue. Unfortunately, thyroid surgery is associated with discomfort, scarring, inconvenience, direct and indirect costs, potential lifelong medication, and occasional surgical complications.[8,9] Efforts to exclude cancer with clinical assessment alone are admittedly imperfect,[5] and laboratory testing of serum thyroid-stimulating hormone levels and thyroid imaging with radionuclides or ultrasonography identify benignity with high confidence in only 4% to 26% of nodules.[10-13] Forty years ago, the application of cytology to thyroid nodule specimens obtained by fine-needle aspiration (FNA) biopsy had a substantial effect on patient management by reducing surgery by one-half and doubling the proportion of cancer among patients who underwent surgery.[3,5] However, approximately one-third of thyroid nodule cytology findings today are cytologically indeterminate,[14,15] with estimated risks of malignancy ranging from 5% to 30%.[16] Consequently, approximately three-quarters of patients with cytologically indeterminate thyroid nodules have been referred for surgery,[17,18] even though 80% ultimately prove to have benign nodules.[15,16,18]

The practice of using preoperative genomic information for thyroid nodule differential diagnosis is more than a decade old, and several commercial and noncommercial genomic approaches are currently available.[19] Performance data from blinded prospective multicenter validation trials are limited and include the gene expression classifier (GEC), in which a machine learning–derived classification algorithm uses messenger RNA transcript expression levels to categorize cytologically indeterminate FNAs as either benign or suspicious.[20] Altered messenger RNA expression can occur for several reasons, including complex upstream interactions that occur because of sequence changes in key core genes or in relevant peripheral genes,[21] the effect of epigenetic changes that occur without DNA sequence alterations, and both internal and external modifiers, such as inflammation and lifestyle or environment.[22,23] In a cohort with a 24% prevalence of malignancy, the GEC accurately identified 90% of malignancies (ie, sensitivity) and 52% of benign nodules (ie, specificity) with indeterminate Bethesda III or IV cytology.[20] It intentionally favored high sensitivity over specificity to ensure the accuracy and safety of a benign genomic result. A test with improved specificity for identification of benign nodules and maintained high sensitivity for malignancy detection could spare even more patients from surgery with an accurate benign genomic result (negative predictive value [NPV]) and increase the cancer yield among those with a suspicious result (positive predictive value [PPV]).

Enhanced technologies for characterizing genomic information, including improved methods for the measurement of RNA transcriptome expression and sequencing of nuclear and mitochondrial RNAs, measurement changes in genomic copy number, including loss of heterozygosity, and the development of enhanced bioinformatics and machine learning strategies, have created the opportunity to develop a new, more robust genomic test. This study describes the blinded clinical validation[24] of the novel genomic sequence classifier (GSC) on a prospective multicenter–derived set of patients with FNA samples whose referral to surgery and histopathological diagnosis were determined in the absence of genomic information.

Key Points

Question What is the performance of a genomic sequencing classifier in cytologically indeterminate thyroid nodules?

Findings In this validation study of 183 patients with 191 cytologically indeterminate thyroid nodules, the genomic sequencing classifier was validated and compared with blinded expert histopathology diagnosis as well as the gene expression classifier. The genomic sequencing classifier had a sensitivity of 91% and a specificity of 68%.

Meaning The genomic sequencing classifier accurately classified more patients with indeterminate thyroid cytology as benign than its predecessor, the gene expression classifier.

Methods

Training and Validation Cohorts

The study was approved by institution-specific institutional review boards as well as by Liberty IRB (DeLand, Florida; now Chesapeake IRB) and Copernicus Group Independent Review Board (Cary, North Carolina). All patients provided written informed consent prior to participating in the study. The training cohort is described in eMethods 1 in the Supplement.

Validation Cohort

Dedicated thyroid nodule FNA specimens and surgical histopathology from nodules 1 cm or larger were collected using a prospective and blinded protocol at 49 academic and community centers in the United States from patients 21 years or older. These samples, stored at −80°C, were previously used to validate the GEC. The details of their enrollment and prespecified inclusion and exclusion criteria have been reported elsewhere.[20] Histopathology diagnoses were previously established by an expert panel of thyroid surgical histopathologists that were blinded to all clinical and molecular data.[20] *BRAF* V600E DNA mutational reference status was established by testing DNA from all samples with the competitive allele-specific TaqMan polymerase chain reaction, as described in eMethods 2 in the Supplement. This independent validation cohort was prespecified and divided into a primary test set comprised of all patients with Bethesda III and IV samples described in the clinical validation of the Afirma GEC[20] with sufficient RNA remaining and a secondary test set comprised of all patients with Bethesda II, V, or VI samples described in the clinical validation of the Afirma GEC[20] with sufficient RNA remaining and not randomly assigned to the training set, as described in eMethods 1 in the Supplement.

Blinding of the Independent Test Set

The following steps were implemented to ensure the independent test set was securely blinded throughout algorithm development and validation (eTable 1 in the Supplement). First, each step was documented in a prespecified protocol and time-stamped on execution. Each team member was assigned a single role and allowed access only to information designated for that role. A randomly generated blinded identification number was assigned to each sample in the validation set by information technology engineers who operated independently of all other teams to ensure that all other personnel were unable to link clinical and genomic data. All historic information that could potentially reveal the clinical label on the independent test set was secured in a password-protected folder prior to the start of algorithm development. Information technology engineers conducted performance testing of the validation test set independently of all other teams. RNA purification, library preparation, next-generation sequencing, RNA sequencing pipeline, feature extraction, and quality control methods are described in eMethods 3-6 in the Supplement.

Algorithm Development

Fine-needle aspiration samples (n = 634) were used to build the GSC core ensemble model, as described in eMethods 1 and eTable 2 in the Supplement. The ensemble model consists of 12 independent classifiers: 6 are elastic net logistic regression models[25] and 6 are support vector machines.[26] The 6 models within each category differ from each other according to the gene sets used (eTable 3 in the Supplement).

To minimize overfitting and to accurately reflect classifier performance incorporating random noise, hyperparameter tuning and model selections were performed using repeated nested cross-validation.[27] Hyperparameter tuning was performed within the inner layer of the cross-validation, and the classifier performance was summarized using the outer layer of the 5-fold cross-validation repeated 40 times. For each classifier, the decision boundary was chosen to optimize specificity, with a minimum requirement of 90% sensitivity to detect malignancy.

The locked ensemble model uses a total of 10 196 genes, among which are 1115 core genes (eTable 4 in the Supplement). These core genes drive the prediction behavior of the model, and the remaining genes improve classifier stability against assay variability.

In addition to the ensemble model described above, the Afirma GSC system includes 7 other components: a parathyroid cassette, a medullary thyroid cancer (MTC) cassette, a *BRAF* V600E cassette, *RET*/*PTC1* and *RET*/*PTC3* fusion detection modules, follicular content index, Hürthle cell index, and Hürthle neoplasm index. The first 4 are upstream of the ensemble classifier, targeting specific and rare patient subgroups (eFigure 1 in the Supplement). The last 3 (the follicular content index, Hürthle cell index, and the Hürthle neoplasm index) were developed to further improve the benign vs suspicious classification performance. They were incorporated with the ensemble classifier to form the core benign vs suspicious classifier engine.

Statistical Analysis

Statistical analyses were performed using R statistical software version 3.2.3 (https://www.r-project.org). Continuous variables were compared using *t* test, and categorical variables were compared using Fisher exact test. We evaluated test performance using sensitivity, specificity, and NPV and PPV based on established methods.[28] All confidence intervals are 2-sided 95% CIs and were computed using the exact binomial test.[29] Test performance comparison between the GSC and GEC was done using McNemar χ^2 test on the matched data set.[30] Significance level in differential gene expression analysis is reported using a false discovery rate–adjusted *P* value.[31] Two-sided *P* values less than .05 were used to declare significance.

Results

We used the FNA samples that previously validated the GEC[20] to independently validate the GSC. The earlier GEC validation samples were derived from 4812 nodule aspirations prospectively collected from 3789 patients at 49 clinical sites in the United States over a 2-year period.[20] Of the 210 validation samples with corresponding Bethesda III or IV cytology and blinded postoperative consensus histopathology diagnoses, 191 (91.0%) had sufficient residual RNA for GSC testing. These samples from cytologically indeterminate nodules constituted the blinded primary test set.

The previously established thyroid nodule cytological diagnosis was used again.[20] Patient demographic characteristics and baseline data are shown in **Table 1**. Age, sex, clinical risk factors, nodule size, histology subtype (eTable 5 in the Supplement), number of FNA passes, prevalence of malignancy (eTable 6 in the Supplement), and proportion of samples collected at community centers did not differ significantly between the primary study population (n = 191) and the GEC clinical validation cohort of samples (n = 210), consistent with unbiased drop out.

The Standards for Reporting of Diagnostic Accuracy Studies was developed to improve the quality of reporting diagnostic accuracy studies.[32] eFigure 2 in the Supplement shows the flow of samples through the study in a Standards for Reporting of Diagnostic Accuracy Studies diagram. Of these 191 indeterminate FNAs, 46 (24.1%) were diagnosed as malignant by an expert surgical histopathology panel who were blinded to all cytologic and genomic results and to the local histopathology diagnosis. Results are reported in the order of testing through the GSC test system (eFigure 1 in the Supplement). Initially, all GSC samples are tested for RNA quantity and quality. None of the 191 samples failed. Subsequently, the GSC aimed to identify nodules composed of parathyroid tissue, those with MTC, and those with a *BRAF* V600E mutation or *RET*/*PTC1* or *RET*/*PTC3* fusion. Samples testing positive for these are included in performance calculations described below, except for samples testing positive for parathyroid tissue, as this result does not indicate a benign or malignant etiology. Among the 191 samples, positive results for parathyroid, MTC, *BRAF*, and *RET*/*PTC* occurred in 0, 1, 3, and 0 samples, respectively. All MTC and *BRAF* V600E results were

Table 1. Baseline Demographic and Clinical Characteristics of the Study Cohort[a]

Variable	GEC Validation	GSC Validation
Total, No.		
Samples	210	191
Patients	199	183
Type of study site, No. (%) of samples		
Academic	76 (36.2)	65 (34.0)
Community	134 (63.8)	126 (66.0)
No. of fine-needle aspiration passes, No. (%) of samples		
1	88 (41.9)	73 (38.2)
2	122 (58.1)	118 (61.8)
Age of patients, mean (range), y	51.2 (22.0-85.0)	51.7 (22.0-85.0)
Sex, No. (%) of patients		
Male	46 (23.1)	41 (22.4)
Female	153 (76.9)	142 (77.6)
Risk factors, No. (%) of patients		
Radiation exposure to head, neck, or both	7 (3.5)	5 (2.7)
Family history of thyroid cancer	14 (7.0)	13 (7.1)
Nodule		
Size on ultrasonography, median (range), cm	2.5 (1.0-9.1)	2.6 (1.0-9.1)
Size group, No. (%) of nodules, cm		
1.00-1.99	69 (32.9)	60 (31.4)
2.00-2.99	62 (29.5)	60 (31.4)
3.00-3.99	42 (20.0)	37 (19.4)
≥4.00	37 (17.6)	34 (17.8)

Abbreviations: GEC, gene expression classifier; GSC, genomic sequencing classifier.

[a] Statistical tests were performed to compare the 191 GSC nodules with the 19 nodules in the GEC validation that were excluded in the GSC validation because of insufficient RNA quantity. The 2 groups differ only on the number of fine-needle aspiration passes, which is not unexpected, as only samples with sufficient remaining RNA were included in the GSC evaluation.

Table 2. Performance of the Genomic Sequencing Classifier (GSC) According to the Final Histopathological Diagnoses and Cytopathological Category

	Reference Standard, % (95% CI)	
GSC Result	Malignant	Benign
Performance across the primary test set of Bethesda III and IV indeterminate nodules (n = 190)		
Suspicious, No./total No.	41/45	46/145
Benign, No./total No.	4/45	99/145
Sensitivity	91.1 (79-98)	
Specificity	68.3 (60-76)	
NPV	96.1 (90-99)	
PPV	47.1 (36-58)	
Prevalence of malignant lesions, %	23.7	
Bethesda III: atypia of undetermined significance/follicular lesion of undetermined significance (n = 114 [60.0%])		
Suspicious, No./total No.	26/28	25/86
Benign, No./total No.	2/28	61/86
Sensitivity	92.9 (76-99)	
Specificity	70.9 (60-80)	
NPV	96.8 (89-100)	
PPV	51.0 (37-65)	
Prevalence of malignant lesions, %	24.6	
Bethesda IV: follicular or Hürthle cell neoplasm or suspicious for follicular neoplasm (n = 76 [40.0%])		
Suspicious, No./total No.	15/17	21/59
Benign, No./total No.	2/17	38/59
Sensitivity	88.2 (64-99)	
Specificity	64.4 (51-76)	
NPV	95.0 (83-99)	
PPV	41.7 (26-59)	
Prevalence of malignant lesions, %	22.4	
Performance across the secondary test set of Bethesda II, V, and VI nodules (n = 61)[a]		
Suspicious, No./total No.	34/34	7/26
Benign, No./total No.	0/34	19/26
Sensitivity	100 (90-100)	
Specificity	73.1 (52-88)	
NPV	100 (82-100)	
PPV	82.9 (68-93)	
Prevalence of malignant lesions, %	56.7	
Bethesda II: cytopathologically benign (n = 19 [31.1%])[a]		
Suspicious, No./total No.	2/2	2/16
Benign, No./total No.	0/2	14/16
Sensitivity	100 (16-100)	
Specificity	87.5 (62-98)	
NPV	100 (77-100)	
PPV	50.0 (7-93)	
Prevalence of malignant lesions, %	11.1	
Bethesda V: suspicious for malignancy (n = 23 [37.7%])		
Suspicious, No./total No.	13/13	5/10
Benign, No./total No.	0/13	5/10
Sensitivity	100 (75-100)	
Specificity	50.0 (19-81)	
NPV	100 (48-100)	
PPV	72.2 (47-90)	

(continued)

concordant with reference methods (eMethods 2 in the Supplement). After this testing, samples were evaluated for follicular cell content by the follicular content index classifier. One sample, negative for the above results, was deemed to have inadequate follicular content and therefore was assigned no result. This sample was excluded from subsequent analyses, leaving 190 samples. **Table 2** summarizes clinical performance characteristics for Bethesda III and IV nodules.

The GSC correctly identified 41 of 45 malignant samples as suspicious, yielding a sensitivity of 91.1% (95% CI, 79-98), and 99 of 145 nonmalignant samples were correctly identified as benign by the GSC, yielding a specificity of 68.3% (95% CI, 6076). Among Bethesda III and IV samples, the NPV was 96.1% (95% CI, 90-99) and the PPV was 47.1% (95% CI, 36-58). Performance of the GSC was similar between Bethesda III and IV categories (Table 2).

Among the 190 Bethesda III and IV samples, 17 (8.9%) were histologically Hürthle cell adenomas and 9 (4.7%) were Hürthle cell carcinomas, while 164 samples (86.3%) were histologically non-Hürthle. For samples with Hürthle histology, the sensitivity was 88.9% (95% CI, 52-100) and the specificity was

Table 2. Performance of the Genomic Sequencing Classifier (GSC) According to the Final Histopathological Diagnoses and Cytopathological Category (continued)

GSC Result	Reference Standard, % (95% CI)	
	Malignant	Benign
Prevalence of malignant lesions, %	56.5	
Bethesda VI: cytopathologically malignant (n = 19 [31.1%])		
Suspicious, No./total No.	19/19	0/0
Benign, No./total No.	0/19	0/0
Sensitivity	100 (82-100)	
PPV	100 (82-100)	
Prevalence of malignant lesions, %	100	

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

[a] One sample has no result because of low follicular content that is not summarized in the table.

Table 3. Performance of Genomic Sequencing Classifier (GSC) According to Histopathological Subtype

Histopathological Subtype	Nodules, No. (%)	Result With GSC, Benign, No./Suspicious, No.
Benign		
Total, No.	145	NA
Benign follicular nodule	49 (33.8)	38/11
Hyperplastic nodule	5 (3.4)	5/0
Follicular adenoma	54 (37.2)	37/17
Follicular tumor of uncertain malignant potential	9 (6.2)	4/5
Well-differentiated tumor of uncertain malignant potential	8 (5.5)	4/4
Hürthle cell adenoma	17 (11.7)	10/7
Chronic lymphocytic thyroiditis	2 (1.4)	1/1
Hyalinizing trabecular adenoma	1 (0.7)	0/1
Malignant		
Total, No.	45	NA
Papillary thyroid carcinoma	15 (33.3)	2/13
Tall-cell variant	1 (2.2)	0/1
Follicular variant	11 (24.4)	1/10
Hürthle cell carcinoma[a]	9 (20.0)	1/8
Follicular carcinoma[b]	7 (15.6)	0/7
Poorly differentiated carcinoma	1 (2.2)	0/1
Medullary thyroid cancer	1 (2.2)	0/1

Abbreviation: NA, not applicable.

[a] Among the Hürthle cell carcinomas, 7 showed capsular invasion and 2 showed vascular invasion. The false-negative case was previously false-negative on the gene expression classifier.[20]

[b] Among the follicular carcinomas, 3 showed capsular invasion and 4 were well-differentiated carcinomas not otherwise specified.

58.8% (95% CI, 33-82). For samples with non-Hürthle histology, the sensitivity was 91.7% (95% CI, 78-98) and the specificity was 69.5% (95% CI, 61-77).

A wide variety of malignant subtypes were correctly classified as suspicious (**Table 3**). Four false-negative cases occurred (**Table 4**). We assessed whether patient age or sex, malignancy subtype, or nodule size by ultrasonography or on histopathology were associated with false-negative cases, and none were. Comparisons of GSC to GEC results on a per-sample basis are reported in the eAppendix in the Supplement. The performance of the GSC in secondary analyses of nodules with Bethesda II, V, or VI cytopathology are reported in Table 2. Among the entire secondary analysis group, the GSC sensitivity was 100% (95% CI, 90-100) and the specificity was 73.1% (95% CI, 52-88).

Discussion

A 2016 meta-analysis[33] reported the risks of malignancy among Bethesda III and IV thyroid nodules to be 17% (95% CI, 11-23) and 25% (95% CI, 20-29), respectively. To safely avoid unnecessary diagnostic surgery among these cytologically indeterminate nodules, a test with a high sensitivity and NPV for malignancy is required. This blinded clinical validation of the GSC in a prospectively collected, representative, universally operated, and histopathologically diagnosed cohort demonstrates the required high NPV across these ranges of cancer prevalence encountered in Bethesda III and IV nodules in clinical practice (**Figure**). To independently validate the GSC, we implemented a set of strict blinding and deidentification protocols that enabled us to use the same FNA samples previously used to validate the GEC.[20] Use of these samples allowed testing of complete and representative sets of nodules with corresponding surgical histology unaffected by the current widespread use of molecular testing to avoid or encourage surgery.

Test sensitivity of the GSC (91%; 95% CI, 79-98) compared with the GEC (89%; 95% CI, 76-96) was maintained, with the point estimate within the counterpart's 95% CI, and the McNemar χ^2 test (df = 1) on the matched sample set renders a test statistic of 0 ($P > .99$). On the other hand, test specificity of the GSC (68%; 95% CI, 60-76) was significantly improved from the GEC (50%; 95% CI, 42-59), with the point estimate outside the counterpart's 95% CI, and the McNemar χ^2 test (df = 1) on the matched sample set renders a test statistic of 16.447 ($P < .001$) (eTable 7 in the Supplement). In practice, this enhanced performance suggests that among Bethesda III and IV nodules that are histopathologically benign, at least one-third more will receive a benign result using the GSC compared with the GEC. At a cancer prevalence of 24%, more than half of tested patients are projected to receive a GSC benign result, and among GSC suspicious nodules, nearly half are anticipated to have cancer on surgical histology. This increased benign call rate is expected to result in more patients being assigned to active observation as opposed to diagnostic surgery. Given the high cost of surgery in the United States among Medicare and private payers,[34] the increased avoidance of diagnostic surgery because of GSC benign results is expected to further improve cost-effectiveness and reduce surgical complications.[8,9]

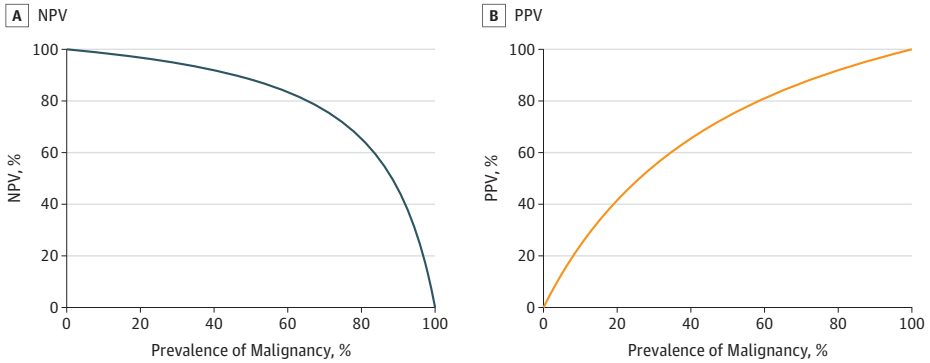
While genomic data has been incorporated in clinical management decisions of multiple medical conditions for more than a decade, progress continues toward understanding the complexities of genomic and nongenomic pathways in the development and behavior of disease. Current evidence suggests that most common diseases are associated with small effects from a large number of genes and that most of these contributions are derived

Table 4. Cytologic Findings and Histopathological Diagnosis in 4 False-Negative Results on Genomic Sequencing Classification

Patient No./Sex	Nodule Size, cm		Bethesda Cytologic Diagnosis	Final Histologic Diagnosis
	Ultrasonographic Imaging	Pathological Examination		
1/M	1.1	1.2	III	PTC
2/F	2.5	1.5	III	PTC
3/F	3.2	3.0	IV	FVPTC
4/F	2.9	3.5	IV	HCC-v

Abbreviations: FVPTC, papillary thyroid cancer follicular variant; HCC-v, Hürthle cell carcinoma, vascular invasion; PTC, papillary thyroid cancer.

Figure. Afirma Genomic Sequencing Classifier Performance Across Differing Risk Populations



There was a negative predictive value (NPV) of 96% (95% CI, 90-99) (A) and a positive predictive value (PPV) of 47% (95% CI, 36-58) (B) at a 24% cancer prevalence in the current Bethesda III and IV cohort. A 2016 meta-analysis[33] reported prevalence of malignancy among Bethesda III and IV nodules as 17% (95% CI, 11-23) and 25% (95% CI, 20-29), respectively. Deriving PPV and NPV at 11% cancer prevalence yielded 98% NPV and 26% PPV, and deriving PPV and NPV at 29% cancer prevalence yielded 95% NPV and 54% PPV.

from transcriptionally active portions of the genome.[21] This implies that diseases such as thyroid cancer are unlikely to be accounted for by the effects of a small number of genes. The fact that few genomic variants are associated with 100% penetrance toward malignant histology suggests that a complex interaction of multiple factors ultimately determines the benign or malignant nature of thyroid nodules.[22,23] As the number of these factors expands, it becomes critical to use machine learning and statistical models to interpret their signals in a trained model to derive an accurate diagnosis.

Hürthle lesions exemplify the challenges inherent in complex biology and the opportunity to harness high-dimensional genomic data for predictive model training and subsequent validation. Most Hürthle cell–dominant Bethesda III and IV thyroid nodules have historically undergone surgery given the potential for Hürthle cell carcinoma, yet most have proven to be histologically benign. The GEC identified these samples at a high NPV, but most were categorized as GEC suspicious.[35] We sought to maintain a high NPV while providing more benign results by including 2 dedicated classifiers to work with the core GSC classifier. Among the 26 Hürthle cell adenomas or Hürthle cell carcinomas reported here, the final GSC sensitivity was 88.9% and the specificity was 58.8%; the GEC sensitivity was 88.9% and the specificity was 11.8% among these same neoplasms. Thus, while the overall GSC sensitivity of 91.1% reported here is comparable with that of the GEC (by design), the improved overall GSC specificity of 68.3% results from significantly improved performances among both Hürthle and non-Hürthle specimen types. Given that most histologically benign Hürthle and non-Hürthle specimens are now both identified as GSC benign, GSC testing may further safely reduce unnecessary surgery among both specimen types.

Recently, the histological diagnosis of noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) was recognized as a biologically distinct entity with a low risk of malignant behavior following surgical excision, which remains the currently recommended treatment.[36] These lesions were previously described as encapsulated noninvasive follicular variant of papillary thyroid cancer.[37] No NIFTP histopathology diagnoses were available in this independent validation cohort, as it was collected prior to the establishment of this diagnostic category. However, subsequent studies[38-40] have suggested a high rate of GEC suspicious results among NIFTP cases. The GSC was trained to identify NIFTP cases as suspicious. While removal of NIFTP from the malignant category would reduce the prevalence of cancers among cytological categories and alter the anticipated PPV of GSC tested cases, this exercise would not be clinically meaningful since the goal of a positive GSC test is to identify all thyroid nodules that warrant surgery, which currently remains necessary for NIFTP.

We performed a secondary analysis of 61 Bethesda II, V, or VI samples that also were included in the GEC validation study (Table 2).[20] While performance of a genomic test among these more definitive cytology categories may not predict performance of the test within the Bethesda III and IV categories, the consistency of these performance metrics is reassuring and supportive of the findings in the primary analysis.

Limitations

Limitations of this study include the lack of performance data among children and data on when the nodule had been previously biopsied or when sample collection methods other than 1 or 2 dedicated FNA passes were used. Another potential limitation is that the prevalence of cancer in this study was toward the higher end of the expected range among Bethesda

III and IV nodules, as seen in the Figure. It is possible that a cytologically indeterminate cohort with a significantly lower prevalence of cancer may contain more benign nodules that are easier for the GSC to classify, as seen in Table 2 among nodules with Bethesda II cytopathology. Should that happen, an effectively higher test specificity may occur.

Conclusions

The current trend in thyroid nodule and cancer management is more conservative, with physicians more aware of the burden of unnecessary thyroid surgery[35] and the indolent behavior of most thyroid malignancies confined to the thyroid.[41-44] Current US guidelines indicate that molecular testing may be used among Bethesda III and IV nodules to add additional information about the nodule's risk of malignancy, which, along with patient preference, may guide clinical decision-making.[7,45] This study demonstrates high test sensitivity and NPV among Bethesda III and IV cytologically indeterminate thyroid nodules across a broad range of nodule sizes (Table 1). As an adjunct to clinical judgment, the GSC is expected to reduce unnecessary diagnostic surgery, improve patient safety, reduce health care costs, and improve patient quality of life.

ARTICLE INFORMATION

Accepted for Publication: February 25, 2018.

Published Online: May 23, 2018. doi:10.1001/jamasurg.2018.1153

Author Affiliations: Division of Endocrine Surgery, Department of Surgery, New York University Langone Medical Center, New York (Patel); Division of Endocrinology, Diabetes, and Hypertension, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts (Angell); Department of Research and Development, Veracyte Inc, San Francisco, California (Babiarz, Huang, Kennedy, Kim, Walsh, Whitney); Department of Medical Affairs, Veracyte Inc, San Francisco, California (Barth, Kloos); Department of Clinical Affairs, Veracyte Inc, San Francisco, California (Barth); Texas Diabetes and Endocrinology, Austin (Blevins); Section of Endocrine Surgery, Department of Surgery, University of California, San Francisco (Duh); Division of Head and Neck Pathology, Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York (Ghossein); The Memorial Center for Integrative Endocrine Surgery, Hollywood, Florida (Harrell); The Memorial Center for Integrative Endocrine Surgery, Weston, Florida (Harrell); The Memorial Center for Integrative Endocrine Surgery, Boca Raton, Florida (Harrell); Anatomic Pathology Division, Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia (LiVolsi); Division of Thyroid and Parathyroid Endocrine Surgery, Department of Otolaryngology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston (Randolph); Head and Neck Pathology Subspecialty, Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston (Sadow); Endocrine Associates of Long Island, Smithtown, New York (Shanik); Section of Endocrine Surgery, Department of Surgery, Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina (Sosa); Thyroid Cytopathology Partners, Austin, Texas (Traweek); Department of Surgery, Endocrine Surgery Program, David Geffen School of Medicine at UCLA, University of California, Los Angeles (Yeh); Division of Endocrinology, Diabetes, and Metabolism, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland (Ladenson).

Author Contributions: Drs Kennedy and Ladenson had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Babiarz, Barth, Harrell, Huang, Kennedy, Kloos, LiVolsi, Randolph, Shanik, Walsh, Whitney, Ladenson.
Acquisition, analysis, or interpretation of data: Patel, Angell, Babiarz, Barth, Blevins, Duh, Ghossein, Huang, Kennedy, Kim, Kloos, Randolph, Sadow, Shanik, Sosa, Traweek, Walsh, Whitney, Yeh, Ladenson.
Drafting of the manuscript: Patel, Babiarz, Barth, Harrell, Huang, Kennedy, Kim, Kloos, Randolph, Shanik, Ladenson.
Critical revision of the manuscript for important intellectual content: Patel, Angell, Babiarz, Barth, Blevins, Duh, Ghossein, Huang, Kennedy, Kim, Kloos, LiVolsi, Randolph, Sadow, Shanik, Sosa, Traweek, Walsh, Whitney, Yeh, Ladenson.
Statistical analysis: Barth, Huang, Kennedy, Kim.
Obtained funding: Kennedy.
Administrative, technical, or material support: Babiarz, Barth, Kennedy, Kloos, Randolph, Sadow, Traweek, Whitney.
Study supervision: Patel, Angell, Barth, Kennedy, Kloos, Randolph, Sosa, Walsh.

Conflict of Interest Disclosures: Drs Patel, Blevins, Shanik, and Ladenson have received speaker's honoraria from Veracyte Inc. Drs Ghossein, LiVolsi, Sadow, and Ladenson serve as consultants for Veracyte Inc. Drs Blevins, Shanik, and Ladenson have received institutional research support from Veracyte Inc. Drs Babiarz, Barth, Huang, Kennedy, Kim, Kloos, and Whitney and Mr Walsh are employees of Veracyte Inc. Drs Babiarz, Barth, Huang, Kennedy, Kim, Kloos, Traweek, and Whitney and Mr Walsh own equity in Veracyte Inc. Dr Sosa is a member of the American Thyroid Association Data Monitoring Committee of the Medullary Thyroid Cancer Consortium, which is supported by GlaxoSmithKline, Novo Nordisk, AstraZeneca, and Eli Lilly. No other disclosures were reported.

Funding/Support: This study was funded by Veracyte Inc.

Role of the Funder/Sponsor: Veracyte Inc drafted the study design and oversaw the data collection, management, and initial analysis. Veracyte Inc had no role in data interpretation; preparation, review, and approval of the manuscript; and the decision to submit the manuscript.

Meeting Presentation: Summary findings from this study were presented as an abstract and oral presentation at the Third World Congress on Thyroid Cancer; July 27-30, 2017; Boston, Massachusetts.

Additional Contributions: We thank the many investigators and patients who provided the fine-needle aspiration samples used here for training and in the independent test set.

REFERENCES

1. Davies L, Welch HG. Current thyroid cancer trends in the United States. *JAMA Otolaryngol Head Neck Surg*. 2014;140(4):317-322.

2. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in thyroid cancer incidence and mortality in the United States, 1974-2013. *JAMA*. 2017;317(13):1338-1348.

3. Mazzaferri EL. Management of a solitary thyroid nodule. *N Engl J Med*. 1993;328(8):553-559.

4. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest*. 2009;39(8):699-706.

5. Hegedüs L. Clinical practice: the thyroid nodule. *N Engl J Med*. 2004;351(17):1764-1771.

6. Kamran SC, Marqusee E, Kim MI, et al. Thyroid nodule size and prediction of cancer. *J Clin Endocrinol Metab*. 2013;98(2):564-570.

7. 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):1-133.

8. Li H, Robinson KA, Anton B, Saldanha IJ, Ladenson PW. Cost-effectiveness of a novel molecular test for cytologically indeterminate thyroid nodules. *J Clin Endocrinol Metab*. 2011;96(11):E1719-E1726.

9. Meltzer C, Klau M, Gurushanthaiah D, et al. Risk of complications after thyroidectomy and parathyroidectomy: a case series with planned chart review. *Otolaryngol Head Neck Surg*. 2016;155(3):391-401.

10. Hong MJ, Na DG, Baek JH, Sung JY, Kim JH. Cytology-ultrasonography risk-stratification scoring system based on fine-needle aspiration cytology and the Korean-Thyroid Imaging Reporting and Data System. *Thyroid*. 2017;27(7):953-959.

11. Virmani V, Hammond I. Sonographic patterns of benign thyroid nodules: verification at our institution. *AJR Am J Roentgenol*. 2011;196(4):891-895.

12. Middleton WD, Teefey SA, Reading CC, et al. Multiinstitutional analysis of thyroid nodule risk stratification using the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol*. 2017;208(6):1331-1341.

13. Tang AL, Falciglia M, Yang H, Mark JR, Steward DL. Validation of American Thyroid Association

ultrasound risk assessment of thyroid nodules selected for ultrasound fine-needle aspiration. *Thyroid*. 2017;27(8):1077-1082.

14. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda System for Reporting Thyroid Cytopathology: a meta-analysis. *Acta Cytol*. 2012;56(4):333-339.

15. Melillo RM, Santoro M. Molecular biomarkers in thyroid FNA samples. *J Clin Endocrinol Metab*. 2012; 97(12):4370-4373.

16. Cibas ES, Ali SZ. The Bethesda System for Reporting Thyroid Cytopathology. *Thyroid*. 2009;19 (11):1159-1165.

17. Cibas ES, Baloch ZW, Fellegara G, et al. A prospective assessment defining the limitations of thyroid nodule pathologic evaluation. *Ann Intern Med*. 2013;159(5):325-332.

18. Wang CC, Friedman L, Kennedy GC, et al. A large multicenter correlation study of thyroid nodule cytopathology and histopathology. *Thyroid*. 2011;21(3):243-251.

19. Onenerk AM, Pusztaszeri MP, Canberk S, Faquin WC. Triage of the indeterminate thyroid aspirate: what are the options for the practicing cytopathologist? *Cancer Cytopathol*. 2017;125(S6): 477-485.

20. Alexander EK, Kennedy GC, Baloch ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. *N Engl J Med*. 2012; 367(8):705-715.

21. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169(7):1177-1186.

22. Herceg Z, Hainaut P. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol Oncol*. 2007;1(1):26-41.

23. Ravegnini G, Sammarini G, Hrelia P, Angelini S. Key genetic and epigenetic mechanisms in chemical carcinogenesis. *Toxicol Sci*. 2015;148(1):2-13.

24. Teutsch SM, Bradley LA, Palomaki GE, et al; EGAPP Working Group. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. *Genet Med*. 2009;11(1):3-14.

25. Friedman J, Hastie T, Tibshirani R, Simon N, Narasimhan B, Qian J. Glmnet: lasso and elastic-net regularized generalized linear models. http://CRAN.R-project.org/package=glmnet. Accessed August 15, 2017.

26. Karatzoglou A, Smola A, Hornik K, Zeileis A. Kernlab: an S4 package for kernel methods in R. *J Stat Softw*. 2004;11(9):1-20. doi:10.18637/jss.v011 .i09

27. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform*. 2014;6(1):10.

28. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ*. 1994;309(6947):102.

29. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404-413. doi:10 .2307/2331986

30. Agresti A. *Categorical Data Analysis*. New York, NY: John Wiley & Sons; 1990.

31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289-300.

32. Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem*. 2015;61(12):1446-1452.

33. Krauss EA, Mahon M, Fede JM, Zhang L. Application of the Bethesda classification for thyroid fine-needle aspiration: institutional experience and meta-analysis. *Arch Pathol Lab Med*. 2016;140(10):1121-1131.

34. Singer J, Hanna JW, Visaria J, Gu T, McCoy M, Kloos RT. Impact of a gene expression classifier on the long-term management of patients with cytologically indeterminate thyroid nodules. *Curr Med Res Opin*. 2016;32(7):1225-1232.

35. Kloos RT. Molecular profiling of thyroid nodules: current role for the Afirma gene expression classifier on clinical decision making. *Mol Imaging Radionucl Ther*. 2017;26(suppl 1):36-49.

36. Strickland KC, Vivero M, Jo VY, et al. Preoperative cytologic diagnosis of noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a prospective analysis. *Thyroid*. 2016;26(10):1466-1471.

37. Nikiforov YE, Seethala RR, Tallini G, 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(8):1023-1029.

38. Wong KS, Angell TE, Strickland KC, et al. Noninvasive follicular variant of papillary thyroid carcinoma and the Afirma gene-expression classifier. *Thyroid*. 2016;26(7):911-915.

39. Jiang XS, Harrison GP, Datto MB. Young investigator challenge: molecular testing in noninvasive follicular thyroid neoplasm with papillary-like nuclear features. *Cancer Cytopathol*. 2016;124(12):893-900.

40. Golding A, Shively D, Bimston DN, Harrell RM. Noninvasive encapsulated follicular variant of papillary thyroid cancer: clinical lessons from a community-based endocrine surgical practice. *Int J Surg Oncol*. 2017;2017:4689465.

41. Davies L, Welch HG. Thyroid cancer survival in the United States: observational data from 1973 to 2005. *Arch Otolaryngol Head Neck Surg*. 2010;136 (5):440-444.

42. Tuttle RM, Fagin JA, Minkowitz G, et al. Natural history and tumor volume kinetics of papillary thyroid cancers during active surveillance. *JAMA Otolaryngol Head Neck Surg*. 2017;143(10):1015-1020.

43. Miyauchi A, Ito Y, Oda H. Insights into the management of papillary microcarcinoma of the thyroid. *Thyroid*. 2018;28(1):23-31.

44. Nou E, Kwong N, Alexander LK, Cibas ES, Marqusee E, Alexander EK. Determination of the optimal time interval for repeat evaluation after a benign thyroid nodule aspiration. *J Clin Endocrinol Metab*. 2014;99(2):510-516.

45. National Comprehensive Cancer Network. Thyroid carcinoma. NCCN Clinical Practice Guidelines in Oncology. Version 2. https://www.nccn.org/professionals/physician_gls /default.aspx. Accessed August 17, 2017.

Invited Commentary

The Problem of the Indeterminate Thyroid Nodule
A Genomic Sequencing Classifier and Clinical Judgment

Peter Angelos, MD, PhD

Despite the value of cytologic evaluation of thyroid nodules in reducing the frequency of surgery for benign cases, the challenge of indeterminate nodules, with their 5% to 30% risk of becoming cancerous, remains.[1,2] Patients with indeterminate thyroid nodules have typically been recommended for surgery even though most nodules proved to be benign.[3] In recent years, a commercially available gene expression classifier (GEC) test with high sensitivity and negative predictive value has been shown to reduce the number of thyroidectomies performed for benign disease.[4]

In this issue of *JAMA Surgery*, Patel et al[5] have used next-generation RNA sequencing and machine learning algorithms to further reduce the numbers of patients who need surgery for nodules that are ultimately shown to be benign. The authors have extracted RNA from a previously collected set of cytologically indeterminate nodules. Through multiple techniques, the RNA transcriptome expression has been used to improve the test compared with a GEC test. This genomic sequence classifier approach has main-

Related article page 817