ORIGINAL ARTICLE

Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology

Erik K. Alexander, M.D., Giulia C. Kennedy, Ph.D., Zubair W. Baloch, M.D., Ph.D., Edmund S. Cibas, M.D., Darya Chudova, Ph.D., James Diggans, Ph.D., Lyssa Friedman, R.N., M.P.A., Richard T. Kloos, M.D., Virginia A. LiVolsi, M.D., Susan J. Mandel, M.D., M.P.H., Stephen S. Raab, M.D., Juan Rosai, M.D., David L. Steward, M.D., P. Sean Walsh, M.P.H., Jonathan I. Wilde, Ph.D., Martha A. Zeiger, M.D., Richard B. Lanman, M.D., and Bryan R. Haugen, M.D.

ABSTRACT

BACKGROUND

Approximately 15 to 30% of thyroid nodules evaluated by means of fine-needle aspiration are not clearly benign or malignant. Patients with cytologically indeterminate nodules are often referred for diagnostic surgery, though most of these nodules prove to be benign. A novel diagnostic test that measures the expression of 167 genes has shown promise in improving preoperative risk assessment.

METHODS

We performed a 19-month, prospective, multicenter validation study involving 49 clinical sites, 3789 patients, and 4812 fine-needle aspirates from thyroid nodules 1 cm or larger that required evaluation. We obtained 577 cytologically indeterminate aspirates, 413 of which had corresponding histopathological specimens from excised lesions. Results of a central, blinded histopathological review served as the reference standard. After inclusion criteria were met, a gene-expression classifier was used to test 265 indeterminate nodules in this analysis, and its performance was assessed.

RESULTS

Of the 265 indeterminate nodules, 85 were malignant. The gene-expression classifier correctly identified 78 of the 85 nodules as suspicious (92% sensitivity; 95% confidence interval [CI], 84 to 97), with a specificity of 52% (95% CI, 44 to 59). The negative predictive values for "atypia (or follicular lesion) of undetermined clinical significance," "follicular neoplasm or lesion suspicious for follicular neoplasm," or "suspicious cytologic findings" were 95%, 94%, and 85%, respectively. Analysis of 7 aspirates with false negative results revealed that 6 had a paucity of thyroid follicular cells, suggesting insufficient sampling of the nodule.

CONCLUSIONS

These data suggest consideration of a more conservative approach for most patients with thyroid nodules that are cytologically indeterminate on fine-needle aspiration and benign according to gene-expression classifier results. (Funded by Veracyte.)

From the Departments of Medicine (E.K.A.) and Pathology (E.S.C.), Brigham and Women's Hospital and Harvard Medical School, Boston; Veracyte, South San Francisco, CA (G.C.K., D.C., J.D., L.F., P.S.W., J.I.W., R.B.L.); the Departments of Pathology (Z.W.B., V.A.L.) and Medicine (S.J.M.), Perelman School of Medicine, University of Pennsylvania, Philadelphia; the Department of Medicine, Ohio State University College of Medicine, Columbus (R.T.K.); the Department of Pathology, University of Washington School of Medicine, Seattle (S.S.R.); Centro Diagnostico Italiano, Milan (J.R.); the Department of Surgery, University of Cincinnati College of Medicine, Cincinnati (D.L.S.); the Department of Surgery, Johns Hopkins University School of Medicine, Baltimore (M.A.Z.); and the Department of Medicine, University of Colorado School of Medicine, Aurora (B.R.H.). Address reprint requests to Dr. Alexander at the Thyroid Unit, Division of Endocrinology, Metabolism and Diabetes, Brigham and Women's Hospital, 75 Francis St., Rm. PBB-B4, Boston, MA 02115, or at ekalexander@partners.org; or to Dr. Kennedy at Veracyte, Inc.,7000 Shoreline Ct., Suite 250, South San Francisco, CA 94080, or at giulia@veracyte.com.

This article was published on June 25, 2012, at NEJM.org.

N Engl J Med 2012.

DOI: 10.1056/NEJMoa1203208

Thyroid nodules are common and are usually benign.[1] However, 5 to 15% prove to be malignant; accordingly, identification of a nodule 1 cm or larger in diameter often prompts a diagnostic evaluation.[2,3] The cornerstone of thyroid-nodule evaluation is fine-needle aspiration,[4] which enables the assessment of cellular morphologic features that could not be identified by means of clinical assessment or imaging. Preoperative, ultrasonographically guided fine-needle aspiration has been shown to accurately classify 62 to 85% of thyroid nodules as benign, thereby avoiding diagnostic surgery.[5]

However, 15 to 30% of aspirations yield indeterminate cytologic findings,[4] which include three subtypes: "atypia (or follicular lesion) of undetermined significance," "follicular neoplasm or suspicious for follicular neoplasm," and "suspicious for malignancy."[6,7] Most patients with cytologically indeterminate nodules are referred for diagnostic thyroid surgery, but the majority prove to have benign disease.[4,8] For these patients, thyroid surgery is unnecessary, yet it exposes them to a 2 to 10% risk of serious surgical complications, and most would require levothyroxine replacement therapy for life.[9-13] These data confirm the critical need to improve the preoperative diagnostic evaluation for patients with indeterminate cytologic findings on fine-needle aspiration.

Molecular analysis of thyroid tissue is poised to become a powerful adjunct to visual microscopical evaluation, since 60 to 70% of thyroid cancers harbor at least one known genetic mutation.[14] Recent investigations have revealed the potential benefits of combined microscopical and molecular analysis of thyroid nodules.[15] When indeterminate aspirates were analyzed for the presence of *BRAF* and *RAS* mutations and for *RET/PTC* and *PAX8–PPARγ* (peroxisome proliferator–activated receptor gamma 1) gene rearrangements, mutations were found in 16% of cases.[16] These genetic markers have high specificity and a high positive predictive value and therefore identify which indeterminate nodules are malignant.[17] Marker positivity can lead to a recommendation for total thyroidectomy rather than for hemithyroidectomy or watchful waiting. With this approach, a second thyroidectomy (so-called completion thyroidectomy) is avoided if the initial hemithyroidectomy reveals malignant nodules.[18] This clinical scenario is similar to that described in reports on the use of epigenetic and peripheral-blood markers.[19,20] Though useful, these markers have limited sensitivity and a limited negative predictive value[21,22] and therefore fail to detect more than 33% of cancers.[16] This rate is too high to be helpful in making the difficult choice between watchful waiting and diagnostic thyroid surgery. Thus, currently available molecular markers fail to rule out cancer with sufficient certainty to avoid surgery in most patients with indeterminate nodules.

Studies have described the development of gene-expression classifiers that better distinguish benign from malignant thyroid nodules.[22] To be of use in avoiding surgery, such a test would need to have high sensitivity and a high negative predictive value. However, these previously reported genomic classifiers remain limited in their sensitivity, and their usefulness has not been validated in large groups of patients.[22] Recently, a gene-expression classifier has been found to help identify nodules that are benign rather than malignant. This classifier was shown to have a sensitivity exceeding 90% and a negative predictive value greater than 95% in a pilot study.[23] We describe the results of a large, prospective, double-blind, multicenter study validating this gene-expression classifier in patients with indeterminate thyroid nodules.

METHODS

STUDY DESIGN AND OVERSIGHT

The study was designed and supervised by the coprincipal academic investigators and by two employees of Veracyte (the makers of this gene-expression classifier), with oversight by a steering committee that met to review the study protocol and analysis. All authors reviewed the study data, vouch for the fidelity of the data and conduct of the study to the protocol, and approved the decision to submit the manuscript for publication. No authors except those who are Veracyte employees serve as consultants or hold equity or equity options in the company. Samples were tested at Veracyte in a laboratory certified according to the provisions of the Clinical Laboratory Improvement Amendments, and the statistical analysis was performed by two authors who are statisticians at Veracyte. The protocol was approved by both central and institution-specific investigational review boards. All patients provided written informed consent to participate in

the study. The coprincipal investigators had full access to all study data and analyses. The first author wrote the first draft of the manuscript, and no one other than the listed authors assisted in the writing.

STUDY POPULATION AND PROTOCOL

We performed a prospective, noninterventional, multicenter validation trial (VERA001) involving patients with ultrasonographically confirmed thyroid nodules, 1 cm or larger in diameter, evaluated by means of routine fine-needle aspiration. Throughout the study, both patients and physicians were unaware of the results of testing with the gene-expression classifier (Afirma). Fine-needle aspiration samples were obtained from patients at 49 U.S. sites (see the Supplementary Appendix, available with the full text of this article at NEJM .org). Study sites were representative of both academic and community centers in 26 states. The fine-needle aspirations consisted of two to five needle insertions within each nodule, and 99% were ultrasonographically guided. Initially, one additional sample was obtained for genomic analysis and was shipped frozen (on dry ice). Midway through the study, however, the protocol was modified: two needle insertions were added to improve the RNA yield, and samples were shipped at a temperature of 2 to 25°C. There was no increase in procedural complications; thus, the commercially available test requires two needle insertions. For each enrolled patient, demographic and thyroid-specific characteristics were recorded. Ultrasonographic data precisely confirmed the location and size of the nodules.

After fine-needle aspiration, local cytologic reports were collected for all patients, and reports without a definitive benign or malignant local diagnosis were reviewed by three expert cytopathologists, who reclassified each report according to three categories of the Bethesda System for Reporting Thyroid Cytopathology: atypia (or follicular lesion) of undetermined significance, follicular neoplasm or lesion suspicious for follicular neoplasm, and lesion suspicious for malignancy.[7] Thyroid surgery was performed on the basis of the clinical judgment of the treating physician at each study site, without knowledge of the results on gene-expression classification. The study was open for enrollment between June 23, 2009, and December 3, 2010, and patients with confirmed surgery scheduled before January 31, 2011, were assessed for eligibility[24,25] (median follow-up time from time of sampling, 301 days). After surgery, local histopathological reports and slides were collected, and biopsied nodules were matched to resected nodules according to size and location. All slides were deidentified and scanned to construct a permanent digital file of microscopical images, and the reference standard diagnosis was determined, as described in the Supplementary Appendix. Results that met the histopathological standard and results from the gene-expression classifier were maintained in two separate, password-protected databases. On completion of the study, unblinding and merging of these data sets were performed by an independent third party. After the results became available, it was determined that 36 samples fell outside the 14-day shipping requirements specified a priori in the protocol, 5 samples did not meet clinical eligibility criteria, and 5 separate fine-needle aspirates represented duplicate aspirations from the same nodules performed at different clinical visits. One additional sample was found to have insufficient referential integrity for inclusion, since the pathology experts could not independently confirm that the ultrasonographically aspirated nodule corresponded to the tissue submitted for histologic analysis. Therefore, 47 samples were excluded from the primary analysis. (The reasons for exclusion are described in detail in the Supplementary Appendix.) Gene-expression data are available at the Gene Expression Omnibus website (www.ncbi.nlm.nih .gov/geo) under accession number GSE34289.

LABORATORY METHODS

Detailed descriptions of the microarray assay, the training of the gene-expression classifier, and the annotation of genes used by the classifier are provided in the Supplementary Appendix and in prior reports.[23] All training samples were independent of the validation set used in this study and comprised thyroid samples with defined diagnoses representing one of two classes: benign or malignant. The locked algorithm uses expression of 167 genes to classify aspirated material from thyroid nodules as either benign or suspicious. There are 142 genes in the main classifier (benign or suspicious) and 25 genes that initially filter out rare neoplasms as the sample is processed through a stepwise diagnostic algorithm in a series of "cassettes." A linear modeling ap-

proach was used for feature selection, and a support-vector machine was used for classification.

STATISTICAL ANALYSIS

Statistical analysis was performed with the use of R software, version 2.13. Continuous variables were analyzed by means of Student's t-test and the Wilcoxon rank-sum test. Categorical variables were analyzed with the use of Fisher's exact test, and the Holm procedure was used to correct for testing associations with multiple clinical variables. Sensitivity, specificity, and negative and positive predictive values were calculated with the use of established methods.[26] Two-sided P values of less than 0.05 were considered to indicate statistical significance. Confidence intervals for proportions are reported as two-sided exact binomial 95% confidence intervals.

RESULTS

CHARACTERISTICS OF PATIENTS AND NODULES

To independently validate the results obtained with the gene-expression classifier, we prospectively collected 4812 nodule aspirates from 3789 patients at 49 clinical sites in the United States over a 19-month period. Of the 4812 samples, 577 were indeterminate (12%); for 413 of the 577 samples, resection was subsequently performed, allowing for blinded histopathological review to define the reference standard. Of the resected samples, 25 were used for training the classifier and for analytic verification studies, and 10 were determined to be ineligible (4 because the patient was younger than 21 years of age, 3 because the samples were not received intact, 2 because the nodule size was <1 cm, and 1 because of a protocol deviation at the time of enrollment); all 35 were excluded from the final analysis.

Using predefined laboratory quality-control metrics, we successfully processed 328 samples through the assay, which resulted in valid classifier results. A standard histopathological diagnosis was available for 312 of these samples (95%). As noted above, 47 samples were excluded, leaving 265 independent, indeterminate nodules for primary analysis (Fig. 1 in the Supplementary Appendix).

Review of local cytopathological reports by the central expert panel confirmed the indeterminate classification of the nodules: 49% of samples were considered to be atypia (or follicular lesions) of undetermined significance, 31% follicular neoplasms or lesions suspicious for follicular neoplasm, and 21% lesions suspicious for malignancy. Baseline characteristics of the patients and nodules are shown in Table 1. Age, sex, clinical risk factors, nodule size, and proportion of samples collected at community versus academic centers did not differ significantly between the primary study population (265 samples) and the entire cohort of patients with resected cytologically indeterminate nodules (413 samples), and these findings are representative of the targeted population.[27] In addition to cytologically indeterminate samples, we evaluated a randomly selected subset of 47 cytologically benign and 55 cytologically malignant surgical samples obtained from an independent group of patients.

PERFORMANCE OF THE GENE-EXPRESSION CLASSIFIER

Figure 1 summarizes the clinical performance characteristics for all relevant sample groups. Of the 265 indeterminate fine-needle aspirates, 85 (32%) were classified as malignant on blinded histopathological review. The gene-expression classifier correctly identified 78 of the 85 malignant samples as "suspicious," yielding a sensitivity of 92% (95% confidence interval [CI], 84 to 97); 93 of 180 nonmalignant samples were correctly identified as benign by the gene-expression classifier, yielding a specificity of 52% (95% CI, 44 to 59). For nodules classified as atypia (or follicular lesions) of undetermined significance, the sensitivity was 90% (95% CI, 74 to 98) and the specificity was 53% (95% CI, 43 to 63). For nodules classified as follicular neoplasms or lesions suspicious for follicular neoplasm, the sensitivity was 90% (95% CI, 68 to 99) and the specificity was 49% (95% CI, 36 to 62). For nodules classified as lesions suspicious for malignancy, the sensitivity was 94% (95% CI, 80 to 99) and the specificity was 52% (95% CI, 30 to 74). The percentage of malignant lesions in these three independent categories was 24%, 25%, and 62%, respectively, yielding respective negative predictive values of 95%, 94%, and 85%. Of 47 samples that were cytologically benign, 3 (6%) were malignant on histopathological review. Although this could be considered an estimate of the false negative rate for cytologically benign samples, the patients may also have undergone surgery on the basis of other characteristics that distinguished

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

Variable	Total Enrollment	Indeterminate Cytologic Findings	Indeterminate Cytologic Findings and Subsequent Surgery	Final Validation Set
Total no.				
Samples	4812	577	413	265
Nodules	4775	567	403	265
Patients	3789	532	378	249
Type of study site — % of samples				
Academic	21.4	34.1	37.3	35.1
Community	78.6	65.9	62.7	64.9
No. of fine-needle aspiration passes — % of samples				
1	54.2	51.5	55.4*	43.4†
2	45.8	48.5	44.6*	56.6†
Age of patients — yr				
Mean	53.2	52.8	51.8*	51.5
Range	18–91	19–85	19–85	22–85
Sex — no. of patients (%)				
Male	696 (18.4)	116 (21.8)	84 (22.2)	55 (22.1)
Female	3093 (81.6)	416 (78.2)	294 (77.8)	194 (77.9)
Risk factors — no. of patients (%)				
Radiation exposure — head, neck, or both	91 (2.4)	14 (2.6)	8 (2.1)	8 (3.2)
Family history of thyroid cancer	174 (4.6)	32 (6)	28 (7.4)*	18 (7.2)
Nodules				
Size on ultrasonography — cm				
Median	1.9	2.2	2.3	2.3
Range	0.6–11	0.75–10.3	0.75–10.3	1–9.1
Size group — no. of nodules (%)				
<1.00 cm	37 (0.8)	4 (0.7)	3 (0.7)	0
1.00–1.99 cm	2503 (52.4)	230 (40.6)	153 (38.0)	102 (38.5)
2.00–2.99 cm	1204 (25.2)	153 (27.0)	111 (27.5)	76 (28.7)
3.00–3.99 cm	621 (13.0)	105 (18.5)	76 (18.9)	45 (17.0)
≥4.00 cm	392 (8.2)	74 (13.1)	60 (14.9)	42 (15.8)
Size not available	18 (0.4)	1 (0.2)	0	0

* P<0.05 for the comparison of results for patients who underwent surgery for indeterminate nodules versus patients who did not.
† P<0.05 for the comparison of results for patients who underwent surgery for indeterminate nodules and were included in the final validation set versus those who were not included.

them from the general population. Nevertheless, the gene-expression classifier correctly identified all 3 of these malignant samples as suspicious. All 55 samples that were cytologically malignant were classified as malignant on histopathological evaluation, and all were considered to be suspicious for malignancy according to the gene-expression classifier (100% sensitivity). A wide variety of malignant subtypes were correctly classified as suspicious for malignancy according to this test (Table 2). These included papillary, medullary, and follicular thyroid carcinomas (including those with oncocytic features); poorly differentiated thyroid carcinomas; and thyroid lymphomas.

There were seven false negative results (Table 3).

Performance across the Primary Data Set of Indeterminate Nodules (N=265)

GEC result	Malignant reference standard (N=85)	Benign reference standard (N=180)
Suspicious	78	87
Benign	7	93

Sensitivity, 92% (84–97); specificity, 52% (44–59); PPV, 47% (40–55); NPV, 93% (86–97); prevalence of malignant lesions, 32%

Atypia of Undetermined Significance or Follicular Lesion of Undetermined Significance (N=129, 48.7%)

GEC result	Malignant reference standard (N=31)	Benign reference standard (N=98)
Suspicious	28	46
Benign	3	52

Sensitivity, 90% (74–98); specificity, 53% (43–63); PPV, 38% (27–50); NPV, 95% (85–99); prevalence of malignant lesions, 24%

Follicular or Hürthle-Cell Neoplasm or Suspicious for Follicular Neoplasm (N=81, 30.6%)

GEC result	Malignant reference standard (N=20)	Benign reference standard (N=61)
Suspicious	18	31
Benign	2	30

Sensitivity, 90% (68–99); specificity, 49% (36–62); PPV, 37% (23–52); NPV, 94% (79–99); prevalence of malignant lesions, 25%

Suspicious for Malignancy (N=55, 20.8%)

GEC result	Malignant reference standard (N=34)	Benign reference standard (N=21)
Suspicious	32	10
Benign	2	11

Sensitivity, 94% (80–99); specificity, 52% (30–74); PPV, 76% (61–88); NPV, 85% (55–98); prevalence of malignant lesions, 62%

Performance on Cytopathologically Benign Samples (N=47)

GEC result	Malignant reference standard (N=3)	Benign reference standard (N=44)
Suspicious	3	13
Benign	0	31

Sensitivity, 100% (29–100); specificity, 70% (55–83); prevalence of malignant lesions, 6%

Performance on Cytopathologically Malignant Samples (N=55)

GEC result	Malignant reference standard (N=55)	Benign reference standard (N=0)
Suspicious	55	0
Benign	0	0

Sensitivity, 100% (93–100); prevalence of malignant lesions, 100%

Performance across the Entire Data Set of Indeterminate Samples, When Both GEC Results and Reference Standard Were Available (before post hoc exclusions) (N=312)

GEC result	Malignant reference standard (N=100)	Benign reference standard (N=212)
Suspicious	87	100
Benign	13	112

Sensitivity, 87% (79–93); specificity, 53% (46–60); PPV, 47% (39–54); NPV, 90% (83–94); prevalence of malignant lesions, 32%

Figure 1. Performance of the Gene-Expression Classifier (GEC), According to the Final Histopathological Diagnoses for Cytologically Indeterminate Samples.

NPV denotes negative predictive value and PPV positive predictive value.

One was a Hürthle-cell carcinoma, and the other six were papillary thyroid carcinomas. To better understand potential causes of false negative results, we measured single molecular markers described in the literature as being elevated in papillary thyroid carcinoma. Two of these markers, cytokeratin 19 and CITED1 (Cpb/p300-interacting transactivator 1) (neither of which was used in the gene-expression classifier), were measured for signal intensity.[28] The expression of both markers was significantly lower in all six papillary carcinoma samples with false negative results than in the samples correctly identified by the classifier; the mean $\log_2$ difference in expression intensity was 1.9 with cytokeratin 19 and 3.0 with CITED1 ($P<0.001$ for both comparisons) (Fig. 2). This finding suggests that assay failure is not responsible for the six false negative cases. We then investigated whether the absence of a papillary thyroid carcinoma signal in the false negative cases could be due to a paucity of thyroid follicular cells in the sample. We evaluated epithelial and thyroid follicular cell content by assaying the following markers: cytokeratin 7, thyrotropin receptor, thyroglobulin, and thyroid transcription factor 1.[29] None of these markers are used by the gene-expression classifier. Expression patterns showed that five of six papillary carcinoma samples with false negative results had low follicular cell content (three samples fell within the lowest 10% of all indeterminate samples, and two other samples within the lowest 20%). For three of the four markers, the difference in follicular content between the samples with false negative results and the samples with true positive results was significant, with a mean $\log_2$ difference of more than 1.4 ($P\leq0.004$ for all comparisons).

Other potential causes of false negative results were considered. The rate of disagreement between two experts on initial blind review was 14% (37 out of 265), and the rate of post-conferral disagreement in defining the reference standard was 2% (see the Supplementary Appendix). However, none of the false negative results were found

Table 2. Performance of Gene-Expression Classifier, According to Histopathological Subtype.

Histopathological Subtype	No. of Nodules (%)	Result with Gene-Expression Classifier
		no. benign/no. suspicious
Benign		
Total	180 (100)	
Benign follicular nodule*	71 (39.4)	41/30
Follicular adenoma	64 (35.6)	37/27
Follicular tumor of uncertain malignant potential	11 (6.1)	5/6
Well-differentiated tumor of uncertain malignant potential	9 (5.0)	4/5
Hürthle-cell adenoma	21 (11.7)	4/17
Chronic lymphocytic thyroiditis	2 (1.1)	0/2
Hyalinizing trabecular adenoma	2 (1.1)	2/0
Malignant		
Total	85 (100)	
Papillary thyroid carcinoma†	42 (49.4)	4/38
Papillary thyroid carcinoma, follicular variant	19 (22.4)	2/17
Hürthle-cell carcinoma‡	10 (11.8)	1/9
Follicular carcinoma§	10 (11.8)	0/10
Medullary thyroid cancer	2 (2.4)	0/2
Malignant lymphoma	2 (2.4)	0/ 2

* One benign follicular nodule was a colloid nodule.
† One papillary thyroid carcinoma was the tall-cell variant.
‡ Among the Hürthle-cell carcinomas, eight showed capsular invasion and two showed vascular invasion.
§ Among the follicular carcinomas, four showed capsular invasion, one showed vascular invasion, four were well-differentiated carcinomas not otherwise specified, and one was a poorly differentiated carcinoma.

among samples for which there was any disagreement. We also tested demographic and clinical factors such as age, sex, ethnic group, radiation exposure, nodule size, and family history. None were associated with false negative results. Analysis of logistic factors, including time from fine-needle aspiration to nucleic acid extraction and time from fine-needle aspiration to surgery, showed no associations. An examination of RNA quality-control metrics, such as RNA integrity, RNA concentration, and microarray quality metrics, also failed to show any association with false negative results. We did notice a trend toward false negative results in smaller nodules, using both ultrasonographic measurements (1.3 cm vs. 2.2 cm, P=0.14) and histopathological measurements (1.2 cm vs. 1.8 cm, P=0.06). In total, these results suggest that insufficient nodule sampling rather than classifier error may be responsible for the false negative results in this study.

DISCUSSION

This study describes the validation of a gene-expression classifier designed to identify benign, rather than malignant, nodules in a large population of fine-needle aspirates with indeterminate cytologic findings. With the use of the gene-expression classifier, the negative predictive value was 95% for aspirates classified as atypia (or follicular lesions) of undetermined significance and 94% for aspirates classified as follicular neoplasms or lesions suspicious for follicular neoplasm, implying that thyroid nodules with these cytologic abnormalities and benign gene-expression classifier results have a post-test probability of malignancy that is similar to the probability for nodules with cytologically benign features on fine-needle aspiration.[5,30] Although the negative predictive value for aspirates with features suspicious for malignancy was lower, at 85%, ascertainment of a 15% risk of

Table 3. Cytologic Findings and Histopathological Diagnosis in Seven Patients with False Negative Results on Gene-Expression Classification.*

Patient No.	Sex	Nodule Size		Cytologic Diagnosis	Pathological Diagnosis		Final Histologic Diagnosis†	RNA	RNA Integrity Number
		On Ultrasonographic Imaging	On Pathological Examination		Expert 1	Expert 2			
		cm						*ng/μl*	
1	Female	2.9	3.5	FN–SFN	HCC	FC	Malignant HCC	8.6	7.7
2	Female	2.2	1.0	SUSP	PTC, follicular variant	PTC	Malignant PTC, follicular variant	31.1	7.6
3	Female	3.2	3.0	FN–SFN	PTC, follicular variant	PTC, follicular variant	Malignant PTC, follicular variant	7.6	7.4
4	Male	1.1	1.2	AUS–FLUS	PTC	PTC	Malignant PTC	6.5	7.2
5	Male	1.3	1.2	AUS–FLUS	PTC	PTC	Malignant PTC	38.2	7.4
6	Female	1.1	0.6	AUS–FLUS	PTC	PTC	Malignant PTC	18.1	6.9
7	Female	1.1	0.6	SUSP	PTC	PTC	Malignant PTC	2.0	6.8

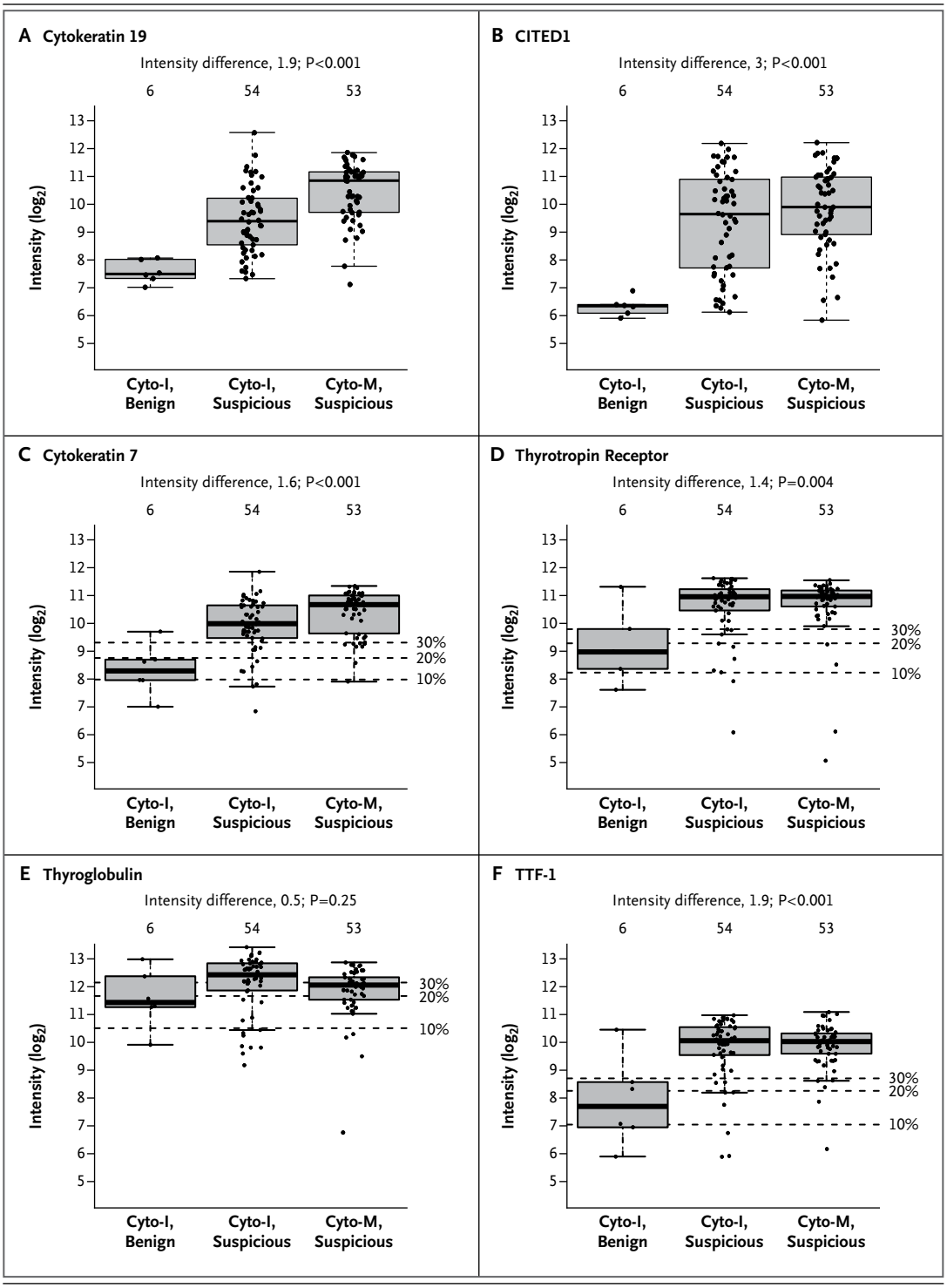
* AUS–FLUS denotes atypia of undetermined significance or follicular lesion of undetermined significance, FC follicular carcinoma, FN–SFN follicular neoplasm or lesion suspicious for follicular neoplasm, HCC Hürthle-cell carcinoma, PTC papillary thyroid carcinoma, and SUSP suspicious for malignancy.
† The final histologic diagnosis was the reference standard.

Figure 2 (facing page). Molecular Signal Intensities in Samples of Papillary Thyroid Carcinoma (Including the Follicular Variant and Tall-Cell Variant).

The box plots indicate the interquartile range and median value; the box plot whiskers indicate the most extreme data points still within 1.5 times the interquartile range from each edge of the box plot. Signal intensity was stratified according to both the cytologic (indeterminate [Cyto-I] or malignant [Cyto-M]) category and the result (or "call") of gene-expression classification (benign or suspicious). In each panel, the box plot on the left shows false negative results (Cyto-I, benign; 6 samples), the box in the center shows true positive results (Cyto-I, suspicious; 54 samples), and the box on the right shows true positive results (Cyto-M, suspicious; 53 samples). Panels A and B show the signal intensity of markers of thyroid cancer (cytokeratin 19 and CITED1, respectively). Panels C through F show the signal intensity of follicular-cell markers (cytokeratin 7, thyrotropin receptor, thyroglobulin, and thyroid transcription factor 1 [TTF-1], respectively). Dashed horizontal lines indicate the 10th, 20th, and 30th percentiles of intensity for the marker in the entire cohort of cytologically indeterminate samples. P values are for the difference in signal intensity between Cyto-I samples classified as benign and those classified as suspicious on the basis of gene-expression classification.

cancer may be useful in deciding whether to perform hemi-thyroidectomy or total thyroidectomy. The observed sensitivity of 100% for cytologically benign and cytologically malignant lesions provides strong independent evidence of the performance of the gene-expression classifier. However, the specificity of 70% for cytologically benign lesions cautions that this test should not be used in the analysis of samples with benign cytologic features. Together, these data suggest that the gene-expression classifier can be useful in making important management decisions, such as recommending watchful waiting in lieu of diagnostic surgery, in the case of nodules with indeterminate cytologic features and benign findings on subsequent testing with the gene-expression classifier.

Patients with well-differentiated thyroid carcinoma have an excellent prognosis, though appropriate surgical management is required.[4,31] Currently, surgery is performed for both diagnostic and therapeutic purposes in patients with indeterminate aspirates. Published reports confirm the high operative efficacy in surgical removal of thyroid cancer, but with a 2 to 10% rate of long-term morbidity from the procedure.[9-11] Thus, surgery should ideally be reserved for therapeutic pur-



poses. The risk of cancer with a benign result on gene-expression classifier testing for nodules classified cytologically as atypia (or follicular lesions) of undetermined significance and for those classified as follicular neoplasms or lesions suspicious for follicular neoplasm is similar to the risk associated with thyroid nodules that have benign cytologic features. Furthermore, implementation of the classifier in routine practice may afford cost savings, with a modest increase in quality-adjust-

ed life-years, primarily by reducing unnecessary surgical resection.[21]

A key strength of this investigation is the inclusion of a wide range of community and academic practice settings, geographic regions, and demographic characteristics of patients. Furthermore, we used local cytopathological reports to classify nodules as having indeterminate cytologic features, and although these reports were reviewed by a central panel of expert cytopathologists to confirm the indeterminate classification, our results reflect test performance based on local cytopathological assessment. This approach makes the findings applicable to everyday patient care. With more than 4000 samples collected, the gene-expression classifier was validated on more than 12 benign and malignant histologic subtypes.[32] Despite the study's strengths, such a protocol also uncovers several immutable realities, creating a practical limit to the test's perfection. For example, even with histopathological analysis by leading experts, independent classifications were discordant in 14% of cases (with a discordance rate of 2% after the experts conferred). Since this served as the reference standard against which the gene-expression classifier was measured, the imperfect interobserver agreement may have affected the sensitivity or specificity of the classifier, since pathological assessment of benign versus malignant disease is not always absolute. Furthermore, five of six false negative results for papillary thyroid carcinoma occurred in samples for which the classifier failed to show independent molecular signatures of papillary thyroid carcinoma and follicular content. This suggests that aspects of nodule sampling such as the technique of fine-needle aspiration or the cellular heterogeneity of the nodule may contribute to inaccurate results.[33] Finally, the prevalence of cancer in the study (32%) may differ from the prevalence previously reported in clinical practice,[3,16] affecting estimates of the negative predictive value. For example, a recent large study showed a 24% prevalence of cancer in clinical practice,[16] and if that prevalence is applied to this study, the overall negative predictive value increases to 95%.

In summary, this study shows that a gene-expression classifier can be used to identify a subpopulation of patients with a low likelihood of cancer in a population of patients for whom diagnostic surgery is otherwise recommended. Though each clinical decision must be individualized, these data suggest consideration of a more conservative clinical approach for patients who have nodules with indeterminate cytologic features on fine-needle aspiration and a benign result on gene-expression classifier testing.

Supported by research grants to the individual study institutions from Veracyte.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin 2012;62:10-29.
2. Gharib H, Papini E, Paschke R, et al. American Association of Clinical Endocrinologists, Associazione Medicie Endocrinologi, and European Thyroid Association medical guidelines for clinical practice for the diagnosis and management of thyroid nodules: executive summary of recommendations. J Endocrinol Invest 2010;33:Suppl:51-6.
3. Yassa L, Cibas ES, Benson CB, et al. Long-term assessment of a multidisciplinary approach to thyroid nodule diagnostic evaluation. Cancer 2007;111:508-16.
4. Cooper DS, Doherty GM, Haugen BR, et al. Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid 2009;19:1167-214. [Erratum, Thyroid 2010;20:674-5.]
5. Wang CC, Friedman L, Kennedy GC, et al. A large multicenter correlation study of thyroid nodule cytopathology and histopathology. Thyroid 2011;21:243-51.
6. Baloch ZW, Cibas ES, Clark DP, et al. The National Cancer Institute thyroid fine needle aspiration state of the science conference: a summation. Cytojournal 2008; 5:6.
7. Cibas ES, Syed AZ. The Bethesda System for reporting thyroid cytopathology. Am J Clin Pathol 2009;132:658-65.
8. Bryson PC, Shores CG, Hart C, et al. Immunohistochemical distinction of follicular thyroid adenomas and follicular carcinomas. Arch Otolaryngol Head Neck Surg 2008;134:581-6.
9. Bergenfelz A, Jansson S, Kristoffersson A, et al. Complications to thyroid surgery: results as reported in a database from a multicenter audit comprising 3,660 patients. Langenbecks Arch Surg 2008;393:667-73.
10. Sosa JA, Bowman HM, Tielsch JM, Powe NR, Gordon TA, Udelsman R. The importance of surgeons experience for clinical and economic outcomes from thyroidectomy. Ann Surg 1998;228:320-30.
11. Shrime MG, Goldstein DP, Seaberg RM, et al. Cost effective management of low-risk papillary thyroid carcinoma. Arch Otolaryngol Head Neck Surg 2007;133:1245-53.
12. Esnaola NF, Cantor SB, Sherman SI, Lee JE, Evans DB. Optimal treatment strategy in patients with papillary thyroid cancer: a decision analysis. Surgery 2001; 130:921-30.
13. Hundahl SA, Cady B, Cunningham MP, et al. Initial results from a prospective cohort study of 5583 cases of thyroid carcinoma treated in the United States during 1996: an American College of Surgeons Commission on Cancer Patient Care Evaluation Study. Cancer 2000;89:202-17.
14. Moses W, Weng J, Sansano I, et al. Molecular testing for somatic mutations improves the accuracy of thyroid fine-needle aspiration biopsy. World J Surg 2010;34:2589-94.
15. Ferraz C, Eszlinger M, Paschke R. Current state and future perspective of molecular diagnosis of fine-needle aspiration biopsy of thyroid nodules. J Clin Endocrinol Metab 2011;96:2016-26.
16. Nikiforov YE, Ohori NP, Hodak SP, et

al. Impact of mutational testing on the diagnosis and management of patients with cytologically indeterminate thyroid nodules: a prospective analysis of 1056 FNA samples. J Clin Endocrinol Metab 2011;96:3390-7.
17. Nikiforov YE, Steward DL, Robinson-Smith TM, et al. Molecular testing for mutations in improving the fine-needle aspiration diagnosis of thyroid nodules. J Clin Endocrinol Metab 2009;94:2092-8.
18. Kim ES, Kim TY, Koh JM, et al. Completion thyroidectomy in patients with thyroid cancer who initially underwent unilateral operation. Clin Endocrinol (Oxf) 2004;61:145-8.
19. Hu S, Ewertz M, Tufano RP, et al. Detection of serum deoxyribonucleic acid methylation markers: a novel diagnostic tool for thyroid cancer. J Clin Endocrinol Metab 2006;91:98-104.
20. Milas M, Shin J, Gupta M, et al. Circulating thyrotropin receptor mRNA as a novel marker of thyroid cancer: clinical applications learned from 1758 samples. Ann Surg 2010;252:643-51.
21. 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.
22. Eszlinger M, Paschke R. Molecular fine-needle aspiration biopsy diagnosis of thyroid nodules by tumor specific mutations and gene expression patterns. Mol Cell Endocrinol 2010;322:29-37.
23. Chudova D, Wilde JI, Wang ET, et al. Molecular classification of thyroid nodules using high-dimensionality genomic data. J Clin Endocrinol Metab 2010;95: 5296-304.
24. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155:529-36.
25. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Am J Clin Pathol 2003; 119:18-22.
26. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;309: 102.
27. Banks ND, Kowalski J, Tsai HL, et al. A diagnostic predictor model for indeterminate or suspicious thyroid FNA samples. Thyroid 2008;18:933-41.
28. Prasad ML, Pellegata NS, Huang Y, Nagaraja HN, de la Chapelle A, Kloos RT. Galectin-3, fibronectin-1, CITED-1, HBME1 and cytokeratin-19 immunohistochemistry is useful for the differential diagnosis of thyroid tumors. Mod Pathol 2005;18:48-57.
29. Fischer S, Asa SL. Application of immunohistochemistry to thyroid neoplasms. Arch Pathol Lab Med 2008;132: 359-72.
30. Lewis CM, Chang K-P, Pitman M, Faquin WC, Randolph GW. Thyroid fine-needle aspiration biopsy: variability in reporting. Thyroid 2009;19:717-23.
31. Bilimoria KY, Bentrem DJ, Ko CY, et al. Extent of surgery affects survival for papillary thyroid cancer. Ann Surg 2007; 246:375-81.
32. Prasad NB, Somervell H, Tufano RP, et al. Identification of genes differentially expressed in benign versus malignant thyroid tumors. Clin Cancer Res 2008;14: 3327-37.
33. Fusco A, Gennaro C, Hui P, et al. Assessment of RET/PTC oncogene activation and clonality in thyroid nodules with incomplete morphological evidence of papillary carcinoma. Am J Pathol 2002;160: 2157-67.