A Large Multicenter Correlation Study of Thyroid Nodule Cytopathology and Histopathology

Chung-Che Charles Wang,[1] Lyssa Friedman,[1] Giulia C. Kennedy,[2] Hui Wang,[2] Electron Kebebew,[3] David L. Steward,[4] Martha A. Zeiger,[5] William H. Westra,[6] Yongchun Wang,[5] Elham Khanafshar,[7] Giovanni Fellegara,[8] Juan Rosai,[8] Virginia LiVolsi,[9] and Richard B. Lanman[1]

Background: Fine-needle aspiration (FNA) biopsies are the cornerstone of preoperative evaluation of thyroid nodules, but FNA diagnostic performance has varied across different studies. In the course of collecting thyroid FNA specimens for the development of a molecular diagnostic test, local cytology and both local and expert panel surgical pathology results were reviewed.

Methods: Prospective FNAs were collected at 21 clinical sites. Banked FNAs were collected from two academic centers. Cytology and corresponding local and expert panel surgical pathology results were compared to each other and to a meta-review of 11 recently published U.S.-based thyroid FNA studies.

Results: FNA diagnostic performance was comparable between the study specimens and the meta-review. Histopathology malignancy rates for prospective clinic FNAs were 34% for cytology indeterminate cases and 98% for cytology malignant cases, comparable to the figures found in the meta-review (34% and 97%, respectively). However, histopathology malignancy rates were higher for cytology benign cases in the prospective clinic FNA subcohort (11%) than in the meta-review (6%, with meta-review rates of 10% at community sites and 2% at academic centers, $p < 0.0001$). Resection rates for prospective clinic FNAs were also comparable to the meta-review for both cytology indeterminate cases (62% vs. 59%, respectively) and cytology malignant cases (82% vs. 81%, respectively). Surgical pathology categorical disagreement (benign vs. malignant diagnosis) was higher between local pathology and a consensus of the two expert panelists (11%) than between the two expert panelists both pre- (8%) and postconferral (3%).

Conclusions: Although recent guidelines for FNA biopsy and interpretation have been published, the rates of false-positive and false-negative results remain a challenge. Two-thirds of cytology indeterminate cases were benign postoperatively and may decrease with the development of an accurate molecular diagnostic test. High disagreement rates between local and expert panel histopathology diagnosis suggests that central review for surgical diagnoses should be used when developing diagnostic tests based on resected thyroid specimens.

Introduction

THE ACCURATE DIAGNOSIS of thyroid nodules continues to challenge physicians managing patients with thyroid disease. The increased use of carotid and other neck ultrasound coupled with the improved technology and higher resolution of ultrasound machines leads to the detection of steadily increasing numbers of asymptomatic thyroid nod-

ules, the so-called incidentalomas (1). Once discovered, these nodules are generally sampled via fine-needle aspiration (FNA) for diagnosis. The rate of thyroid nodule FNA biopsies increased threefold from 1995 to 2005 (2).

Another part of the increase in thyroid nodule FNAs may also reflect growing awareness that the incidence of thyroid cancer is rising (3,4). Increased cancer detection has occurred in small and large malignant nodules alike over the last

decade, suggesting that the rise in cancer incidence is not solely a function of increased diagnostic scrutiny with ultrasound (5–7).

Historically, only 5% of thyroid nodule FNA biopsies were malignant (8), but a recent large retrospective study conducted at a high-volume academic center using ultrasound-guided FNA (UGFNA) on thyroid nodules larger than 1 cm diameter found that about 10% of FNAs proved malignant postoperatively. Half of the malignant nodules were diagnosed by cytology as malignant, and the other half had indeterminate cytology, a term used to describe atypia of undetermined significance (atypia), follicular neoplasm, and suspicious for malignancy (9). In this study, 22% of all FNA biopsies were indeterminate. These findings were consistent with a recent review by Lewis et al. of 20 large (>200 patient) thyroid FNA series published between 2001 and 2006, which found that a median of 24% of FNA biopsies had indeterminate cytology (10). The primary challenge of thyroid nodules with indeterminate cytology is that while most are surgically resected, the majority are found to be benign postoperatively; follicular neoplasms comprise the largest cytopathologic group with malignancies found only about 20% of the time (11–13).

Additionally, the Lewis review (10) found a postoperative risk of malignancy in nodules with benign cytology of 7% (negative predictive value [NPV] 93%), similar to the risk of malignancy for benign nodules at the authors' own institution of 8% (NPV 92%). These rates of postoperative malignancy on nodules with benign cytology diagnoses are similar to those reported in the 2009 American Thyroid Association (ATA) guidelines of 5% (NPV 95%) (14). Variability in thyroid FNA diagnostic test performance was the key finding in the Lewis study (10), with the positive predictive value (PPV) ranging from 15.8% to 74.8% and the NPV ranging from 74% to 98.2%.

Given the variability in performance of FNA diagnosis, and especially considering the high rates of benign thyroid nodules postsurgery in the cytologically indeterminate nodules, guidelines have been written to improve standardization and technique of both FNA collection and interpretation of results (15). Because of these challenges, we have prospectively collected FNA specimens from thyroid nodules as part of a large, multicenter discovery effort to develop a novel molecular diagnostic test (16). The purpose of this test is to better identify the benign nodules with cytologically indeterminate diagnoses preoperatively on FNA samples, so that watchful waiting can be employed in lieu of surgical resection of the thyroid.

In collecting these FNA specimens and analyzing the associated clinical data, our aim is threefold: (i) to correlate the cytology and surgical pathology data to review FNA diagnostic performance (sensitivity, specificity, PPV, NPV) in a prospective study; (ii) to perform an updated meta-review of large observational FNA studies published in the United States from 2002 to 2010 and compare these results to the diagnostic performance results of the prospective FNA study specimens; and (iii) to utilize a panel of outside experts to perform surgical pathology review of cases and evaluate local-to-expert panel pathology concordance.

Materials and Methods

Study specimen collection methodology

From August 2008 through January 2010, FNA specimens and their associated clinical data were collected prospectively from 16 U.S. community-based clinics, 3 U.S. academic centers, and 2 non-U.S. academic sites. For prospective FNA collection, patients were enrolled in an institutional review board-approved protocol and informed consent was obtained. Prospectively collected specimens were collected in clinic, preoperatively, or ex vivo after surgical resection. Retrospectively collected banked FNA specimens were also obtained from two academic centers during this timeframe. The banked FNA specimens were collected in clinic preoperatively at one site and intraoperatively (after surgical dissection has begun and the nodule can be observed), at the other. Cytopathology slides from FNA specimens and histopathology slides from resected thyroid tissues were prepared in accordance with the local standard.

Study specimen clinical data procurement and classification methodology

Age, gender, cytopathology diagnosis, and cytopathology report were obtained for each specimen when available. Each cytopathology result was reviewed and adjudicated by a subset of the authors according to the Bethesda System for Reporting Thyroid Cytopathology (15). According to the Bethesda criteria, the cytopathologic diagnosis of thyroid FNAs falls into six categories: (i) benign (Cyto B); (ii) atypia of undetermined significance or follicular lesion of undetermined significance (ATYP); (iii) follicular neoplasm or suspicious for follicular neoplasm and Hürthle cell neoplasm or suspicious for Hürthle cell neoplasm (FoN/HN); (iv) suspicious for malignancy (SUSP M); (v) malignant (Cyto M), and (vi) nondiagnostic or unsatisfactory (Cyto ND). The ATYP, FoN/HN, and SUSP M diagnostic groups were grouped into a single cytologically "indeterminate" (Cyto I) category, since not all cytopathologists at each clinical site have adopted the Bethesda System and, therefore, it could not be consistently determined whether a case was ATYP, FoN/HN, or SUSP M. This effectively created four major cytology diagnostic categories: Cyto B, Cyto I, Cyto M, and Cyto ND.

For prospectively collected FNAs, sites were contacted monthly for patient follow-up to determine whether surgical resection had occurred. If so, surgical pathology diagnosis and corresponding surgical pathology reports and histology slides were obtained, if available and procurable. For banked FNAs, surgical diagnoses were obtained for each specimen, and corresponding surgical pathology reports and histopathology slides were procured if available. Local surgical diagnoses were reviewed and adjudicated by a subset of the authors and listed according to the World Health Organization criteria (17). When surgical pathology reports diagnosed the nodule of interest as benign yet found incidental papillary carcinomas <1 cm in diameter (microcarcinomas), those nodule diagnoses were classified benign, whereas microcarcinomas in the nodule of interest and not clearly incidental were classified malignant.

All available histopathology slides were sent to two expert pathologists for central review. The two pathologists were blinded to the local and each others' pathology diagnoses. If the experts did not come to complete agreement on their independent, blinded review, they were unblinded to each others' diagnosis, conferred on the case, and came to a consensus diagnosis. Expert panel diagnoses were listed according to the World Health Organization criteria (17), with

the addition of the recommendations from the Chernobyl Pathology Group (18), which includes the use of diagnostic category uncertain malignant potential. Histological diagnoses were classified categorically as either benign or malignant. Minimally invasive follicular neoplasms were considered malignant, whereas well-differentiated neoplasms without capsular or vascular invasion, or definite nuclear changes, were considered benign. This latter category included the diagnosis of uncertain malignant potential, which was used by the expert panel but not the local pathologists.

In addition to collecting agreement data between local and pathologists on the expert panel, as well as between the two expert panelists, Cohen's kappa statistics (chance adjustment of *kappa statistics*) (19) were used to assess the degree of agreement between the pathologists. Kappa values were reported for local-to-expert panel and expert-to-expert comparisons, including 95% confidence intervals (CIs).

Updated meta-review methodology

For the updated meta-review, U.S.-based thyroid FNA biopsy (FNAB) series published between 2002 and 2010 were identified using the PubMed search engine of the National Library of Medicine and National Institutes of Health with appropriate search terms. Criteria for study inclusion were being U.S.-based, utilization of UGFNA in thyroid nodules too small to be aspirated by palpation, and >150 resected specimens with availability of corresponding histopathological diagnoses. The studies that were identified and met the inclusion criteria either reported on all FNAB from a specific time period with histopathological correlation only for those cases proceeding to surgery or reported on surgical cases from a specific time period with cytopathological correlation. Both scenarios utilize the same statistical approach in comparing FNAB cytopathological to histopathological data. Studies were considered "academic" if they were part of an Association of American Medical Colleges–accredited medical school and "community" if they were not.

For the updated meta-review, the surgically resected percentage for each cytopathology subtype was calculated as the number of cases resected in a cytological subtype divided by the total FNAs in that subtype. The histopathology malignant percentage from each cytopathology subtype was calculated as histopathology malignant cases in a cytological subtype divided by the total number of cases resected for each cytological subtype. In deriving the overall surgically resected percentage and the histopathology malignant percentage for the entire 11-study meta-review, the cases for all studies where these calculations could be performed were included and the numbers were pooled, resulting in an overall average among all eligible studies.

FNA diagnostic performance methodology

FNA diagnostic performance was defined by sensitivity, specificity, PPV, and NPV. Indeterminate and malignant FNAs were considered positive test results, as these lead to a clinical recommendation of surgical management. Cyto B FNAs were considered negative test results, and Cyto ND FNAs were excluded from statistical analyses of cytological test performance. True-positives (TP) were defined as nodules with indeterminate or malignant cytology and a corresponding malignant postoperative histology result.

TABLE 1. NUMBER OF FINE-NEEDLE ASPIRATIONS AND SURGERIES FOR PROSPECTIVELY COLLECTED CLINIC FINE-NEEDLE ASPIRATION STUDY SPECIMENS

	Local cytology subtype [no. (%)]				
	B	I	M	ND	Total
No. of FNAs (%)	605 (80)	61 (8)	51 (7)	36 (5)	753
No. of surgeries in each cytology subtype (%)	27 (4)	38 (62)	42 (82)	5 (14)	112

B, benign; FNAs, fine-needle aspirations; I, indeterminate; M, malignant; ND, nondiagnostic.

True-negatives (TN) had both benign cytology and histology. False-negatives (FN) were defined as nodules with benign cytology and malignant histology. False-positives (FP) had indeterminate or malignant cytology and benign histology. The following formulas were employed: sensitivity = [TP/(TP+FN)]; specificity = [TN/(TN+FP)]; PPV = [TP/(TP+FP)]; NPV = [TN = (TN+FN)]. Additionally, the postoperative risk of malignancy on a cytologically benign nodule was defined as 1 − NPV, that is, the percentage of all benign cytological diagnoses that were false-negatives, as the latter expression is more relevant than the false-negative rate to clinical management. The cytological test performance for FNAs was calculated for both the specimens in the molecular discovery study and for the updated meta-review.

## Results

Study specimen results

A total of 1501 FNA specimens were collected from 1285 patients; 606 FNA specimens had surgical pathology results, of which 221 were evaluated by a panel of two expert pathologists. In the clinic, 753 FNAs from 613 patients were collected prospectively. The average patient age was 52 years (range 18–94 years), and 85% of patients were women.

Table 1 lists the cytology diagnoses of all 753 prospectively collected FNA specimens, as well as the number and percentage that went to surgery within each cytology subtype.

TABLE 2. LOCAL SURGICAL HISTOLOGY RESULTS BY LOCAL CYTOLOGY SUBTYPE FOR PROSPECTIVE CLINIC FINE-NEEDLE ASPIRATIONS

	Cytology subtype [no. (%)]	
Surgical result	Indeterminate	Malignant
Benign nodule	1 (3)	
Colloid nodule		1 (2)
Follicular adenoma	9 (24)	
Hürthle cell adenoma	2 (5)	
Lymphocytic thyroiditis	3 (8)	
Nodular hyperplasia	10 (26)	
Follicular carcinoma	1 (3)	
PTC, follicular variant	6 (16)	7 (17)
Micro PTC	2 (5)	7 (17)
PTC	4 (10)	27 (64)

Incidental microcarcinomas were considered benign.
PTC, papillary thyroid carcinoma.
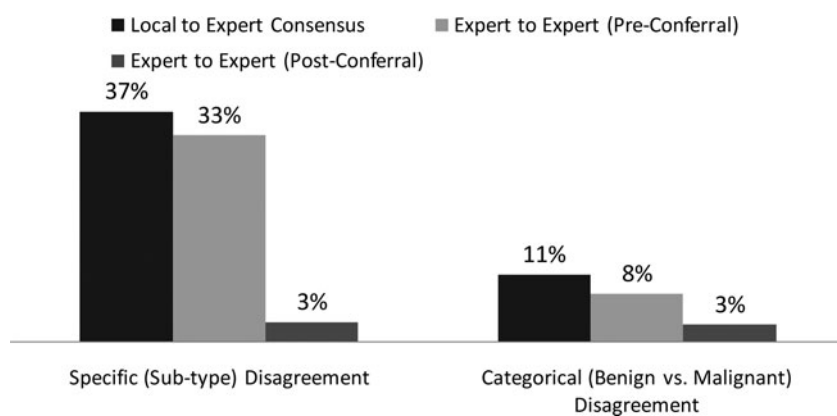
FIG. 1. Categorical disagreement defined as a benign to malignant mismatch (e.g., papillary thyroid carcinoma [PTC] vs. hyperplastic nodule, follicular carcinoma [FC] vs. follicular adenoma [FA], etc.). Specific (subtype) disagreement defined as any mismatch in diagnosis (e.g., FA vs. hyperplastic nodule, PTC vs. follicular variant [FV] PTC, etc., includes categorical disagreements). $n = 221$ surgical pathology cases read by both expert panelists.

The average indeterminate rate for the prospectively collected clinic specimens was 8%, but clinic-specific indeterminate rates ranged from 0% to 40%. After excluding clinical sites that had fewer than 20 specimens, the average indeterminate rate was 7% (range 0%–21%, standard deviation 6.7%).

Of the 112 clinic FNA specimens from nodules that were surgically resected, the malignancy rates for each cytology subtype were as follows: Cyto B 11% (3, 11% micro-carcinoma); Cyto I 34% (13, 29% malignant, including 5% microcarcinoma); Cyto M 98% (41, 81% malignant, including 17% microcarcinoma); Cyto ND 0%. Table 2 shows the specific local surgical histology results for the indeterminate and cytology malignant subtypes.

Of all specimen types (i.e., prospectively collected in the clinic, pre- and postoperative, and banked), slides from 221 resected thyroid nodules were available and reviewed by two expert pathologists (see Fig. 1). Results of each expert were compared against local pathology as well as with each other. Agreement on the specific subtype diagnosis (e.g., follicular adenoma or papillary thyroid cancer) was as follows: 56% local-to-expert1, 59% local-to-expert2, and 67% expert1-to-expert2. Categorical (i.e., benign vs. malignant) disagreements were 8% between expert1 and expert2 (observed agreement 92%, kappa = 0.84, CI 0.77–0.90), less than the 10% disagreement rate between local and expert1 (observed agreement 90%, kappa = 0.79, CI 0.69–0.86), and 13% disagreement rate between local and expert2 (87%, kappa = 0.75, CI 0.65–0.82). When the two expert panel pathologists did not agree on subtype diagnosis and subsequently conferred, their exact subtype match rate increased from 67% to 97% (214 out of 221 specimens) and categorical benign versus malignant agreement increased from 92% to 97%. Local pathology compared to the expert panelists' consensus diagnoses had an

11% benign versus malignant disagreement rate (observed agreement 89%, kappa = 0.78, CI 0.69–0.86).

Updated meta-review results

Eleven U.S.-based thyroid FNA series were identified that met the inclusion criteria between 2002 and 2010 (9,20–29). Of the studies that gave statistics for age and gender, average age was 51 years (46–56) with 14,551 (85%, range 75%–88%) women and 2508 (15%, range 12%–25%) men. Table 3 shows that FNA diagnostic performance (based on sensitivity, specificity, PPV, and NPV) for the meta-review was comparable to the performance of the prospective clinic FNA specimens. Table 4 shows all studies in the updated meta-review and the number of cases with surgical resection and histopathology malignant percentage for each subtype. Additionally, Table 5 shows a side-by-side comparison between the updated meta-review and the prospective clinical FNA collection study data for overall cytology subtype, postoperative diagnosis of malignancy of each cytology subtype, and resection rate of each cytology subtype. The table illustrates the comparable histopathology malignancy rates and resection rates of both the cytology indeterminate and cytology malignant subtypes.

Table 6 describes the percentage of postoperative diagnosis of malignancy for each cytology subtype as well as the false-negative percentage (i.e., cytology benign but histology malignant) for the academic and community sites. A comparison between community and academic sites of postsurgical malignant diagnoses showed a highly statistically significant difference. The pooled average of postsurgical malignant diagnoses for nodules with benign cytology was 10% for community sites and 2% for academic sites ($p < 0.0001$).

Table 3. Fine-Needle Aspiration Diagnostic Performance

Study	Sensitivity TP/(TP+FN)	Specificity TN/(TN+FP)	PPV TP/(TP+FP)	NPV TN/(TN+FN)
Updated meta-review	95%	47%	52%	94%
Prospective clinic FNA specimens only	95%	48%	68%	89%

FN, false–negative; FP, false-positive; NPV, negative predictive value; PPV, positive predictive value; TN, true-negative; TP, true-positive.

TABLE 4. POSTOPERATIVE MALIGNANCY RATES FOR EACH CYTOLOGY SUBTYPE IN UPDATED META-REVIEW

Study	No. resected in series	% malignant postoperatively by cytopathology diagnosis						
		B	ATYP	FoN/HN	SUSP M	I	M	ND
Blansfield et al. (20) (Abington Hospital, PA)	183	18	0	31	57	33	93	0
Sclabas et al. (21) (MD Anderson)	240	4	N/A	16	82	27	96	9
Castro and Gharib (22) (Mayo Clinic)	1598	N/A	N/A	14	65	N/A	N/A	N/A
Wu et al. (23) (Ball Memorial, IN)	381	8	48	26	68	34	100	12
Yassa et al. (9) (Brigham & Women's)	1242	2	24	28	60	42	97	10
Yang et al. (24) (Northshore LIJ & U. Texas Galveston	1052	7	19	32	65	38	99	11
Oertel et al. (25) (Washington Hospital)	1287	10	N/A	49	42	48	95	50
Banks et al. (26) (Johns Hopkins)	639	N/A	33	30	62	37	N/A	N/A
Nayar and Ivanovic (27) (Northwestern)	1413	2	6	15	53	14	96	9
Theoharis et al. (28) (Yale)	378	5	48	34	87	47	100	32
Faquin and Baloch (29) (MGH & University Pennsylvania)	524	N/A	19	25	N/A	N/A	N/A	N/A
Summary	8937	6	16	25	62	34	97	12

ATYP, FN/HN, and SUSP M combined are Indeterminate.

ATYP, atypia of undetermined significance; FoN/HN, follicular/Hürthle cell neoplasm; SUSP M, suspicious for malignancy; I, indeterminate; M, malignant; ND, nondiagnostic; N/A, not applicable.

Discussion

In the course of development of a thyroid FNA molecular diagnostic test, we have collected 1501 specimens of which 753 constituted the largest multicenter prospectively collected thyroid FNA cohort evaluating the cytological and histopathological correlation of thyroid nodules. Of note, 221 FNA specimens had corresponding surgical pathology review by a panel of two external pathology experts blinded to the original diagnosis, with the goal of determining a final adjudicated gold-standard diagnosis, defined as the expert panel consensus diagnosis used to train and validate the molecular diagnostic test. Bartolazzi et al. conducted a large (294 FNA) prospective multicenter study of atypical and follicular neoplasm (Thy3) (30) lesions with external expert histopathological diagnosis, but this study did not include biopsies cytologically suspicious for malignancy (Thy4), nor biopsies with benign or malignant cytological diagnoses, and correlation between local cytological and histopathological diagnosis was not reported (31). Theoharis et al. collected 3207 FNAs from 2468 patients prospectively in 2008, but this was from a single academic center (28).

Of the 753 prospectively collected clinic FNA specimens, 80% were Cyto B, 8% Cyto I, 7% Cyto M, and 5% Cyto ND. These percentages were comparable to the updated meta-review of 11 large U.S.-based FNA series with corresponding cytopathology and surgical pathology results (72% Cyto B, 17% Cyto I, 5% Cyto M, and 6% Cyto ND), except for the lower rates of cytologically indeterminate specimens in the prospective study (Table 5). Because clinical study sites were reimbursed for obtaining FNA specimens of any cytology, the low indeterminate rate may have resulted from an increased recruitment of FNAs with benign cytology, and may not be representative of actual clinical practice.

Our prospectively collected clinic FNA data also demonstrated a wide range of cytology indeterminate diagnostic rates across the various study sites, consistent with the key finding of Lewis' review (10); that is, there is high variability in classifying FNA results into the indeterminate category. Of note, the National Cancer Institute Consensus Conference of thyroid FNA cytology reported that some cytopathologists indeterminate rates as low as 6%, whereas others have indeterminate rates as high as 30% (32). This variability may challenge approaches using even finer diagnostic distinctions, such as the six-category Bethesda classification system. Future studies incorporating blind central review by expert cytopathologists should be considered for quality review of the cytopathology diagnoses.

In the prospectively collected clinic FNA cohort, the postoperative malignancy rates of 34% (29% malignant and 5% microcarcinoma) for Cyto I specimens and 98% (81% malignant and 17% microcarcinoma) for Cyto M specimens corresponded with the updated meta-review findings of 34% malignancy rate for Cyto I nodules and 97% malignancy rate for Cyto M nodules (Table 5). Also, subjects enrolled in the prospective series did not differ in age and gender from the overall values for the meta-review results. These data from the prospective clinic FNA cohort of the study were surprising in that the utilization of UGFNA did not seem to improve the postoperative rate of malignancy in cytologically indeterminate samples compared to the studies in the meta-review, the latter relying in general on combinations of UGFNA and palpation-guided FNA (PGFNA). In addition, even though our prospective FNA collection occurred very recently (between 2008 and 2010), the presumed incorporation of newer techniques and standards from guidelines did not improve FNA diagnostic performance compared to the meta-review data.

Postoperative malignancy rate in benign cytology FNAs

Of the specimens that were diagnosed as cytologically benign and that underwent surgery, 11% were malignant. Although the malignant specimens were microcarcinomas (i.e., microscopic papillary carcinomas), they were the nodules sampled by FNA, and thus were not incidental. These findings represent a higher postoperative malignancy rate than that reported in the 2009 ATA guidelines (14), which was 5% and by the earlier Lewis meta-review (10) of 7%. This

TABLE 5. COMPARISON BETWEEN META-REVIEW AND PROSPECTIVE CLINICAL FINE-NEEDLE ASPIRATION STUDY DATA

	B		I		M		ND	
	Meta-review	Prospective clinical FNA study data	Meta-review	Prospective clinical FNA study data	Meta-review	Prospective clinical FNA study data	Meta-review	Prospective clinical FNA study data
Overall cytology subtype	72% (62%–85%)	80%	17% (10%–26%)	8%	5% (1%–8%)	7%	6% (1%–11%)	5%
Postoperative diagnosis of malignancy	6% (2%–18%)	11%	34% (14%–48%)	34%	97% (93%–100%)	98%	12% (0%–50%)	0%
Resection rate	9% (3%–16%)	4%	59% (48%–81%)	62%	81% (57%–90%)	82%	15% (5%–29%)	14%

could be the result of too low of a threshold for making a benign cytological diagnosis in the prospective study, as reflected by the 80% rate of benign cytology diagnoses and the low rate of indeterminate cytology diagnoses. Additionally, treatment selection bias may play a role, as in the absence of other clinical risk factors most patients with benign FNA cytology do not undergo surgery. Findings may also be secondary to small sample size in the prospective cohort, as only 4% of benign FNAs were operated upon. However, because some of the series in the meta-review were composed of a mix of PGFNA and UGFNA, and the current study was 99% UGFNA, we had expected less sampling error and therefore a lower false-negative rate on cytologically benign nodules in the current study. Ultrasound guidance should have reduced the risk of the FNA missing a malignant nodule and leading to an erroneous benign cytology result in the prospective clinic FNA cohort, but this was not the case given the 6% and 11% risks of malignancy in the meta-review and the prospective clinic FNA collection, respectively. Yeh *et al.*, in a 2004 series of 100 consecutive resected thyroid nodules with benign cytology, found 21% to be malignant postoperatively; the authors cautioned that benign cytology results should be considered in the context of the total clinical presentation and followed diligently (33).

The updated meta-review found significant variation in the risk of postoperative malignancy on nodules with benign cytology (range 2%–18%; Table 5). There was a weighted average of 10% false-negatives on benign nodules in the community as opposed to 2% in academic centers (Table 6), which was a statistically significant difference ($p < 0.0001$). Although the 11 studies were published relatively recently, only one was 100% UGFNA, making it impossible to evaluate whether false-negatives in some studies resulted from higher sampling error with PGFNA. False-negatives could also be a function of quality of FNA sampling, cytological interpretation differences between community and academic sites, or both. Our findings of higher rates of false-negatives in community-based practices versus academic practices are consistent with a report from Norway, which also found significant differences between an academic center and two community sites (34).

Our prospectively collected clinic FNA study specimens were 99% UGFNAs, yet the postoperative risk of malignancy for cytology benign cases was fairly high (11%) in a sample set where 72% of specimens were from nonacademic sites. Although this high percentage of false-negative results could be a spurious finding related to small sample size, our multicenter FNA collection study data, updated meta-review, and Lewis' previously published review (10) all indicate that the percentage of false-negatives (i.e., cytology benign that were postoperatively malignant) for cytologically benign yet resected thyroid FNAs are at least 6%–7%, if not higher.

Microcarcinomas

Microcarcinomas comprised a small but noteworthy percentage of the malignant surgical diagnoses in our overall specimen collection. As papillary thyroid microcarcinoma (mPTC) rises as a percentage of all cancers, as described by Elisei *et al.* in which microcarcinomas rose from 8% before 1990 to 29% from 1990 to 2004 (35), clinicians are increasingly challenged as to which of these relatively indolent micro-

Table 6. Postoperative Diagnosis of Malignancy Based on Cytology Results for Meta-Review

Preoperative diagnosis by cytopathology	B	I	M	ND
Postoperative diagnosis of malignancy (range)	6% (2%–18%)	34% (14%–48%)	97% (93%–100%)	12% (0%–50%)
False-negative rate, academic sites	2%–5%	N/A	N/A	N/A
False-negative rate, community sites	7%–18%	N/A	N/A	N/A

carcinomas may remain unresected and followed clinically. In a nonrandomized case–control study, Ito *et al.* followed 340 patients with mPTC who did not opt for surgical resection and who did not have higher risk clinical features, such as lateral lymph node enlargement or highly undifferentiated cytological features, for an average of 74 months (range 18–187 months), and reported that only 1.4% developed novel nodal metastasis in 5 years and 3.4% in 10 years (36). If most mPTCs behave clinically like benign lesions, and the micro-carcinomas are excluded from the current study, then the true malignancy rates for cytology benign nodules would be lower than reported here, although there would be an even higher rate of benign nodules within the cytology indeterminate category than is estimated here.

Surgical resection rates of cytology indeterminate and malignant FNAs

For our prospectively collected clinic FNA specimens, surgical resection rates were 62% for indeterminate specimens and 82% for cytology malignant specimens. These results were lower than expected based on ATA guidelines, and unlike most published FNA series, physician investigators were contacted monthly to monitor for surgery for up to 1 year. In fact, these figures corresponded closely to the resection rates in the updated meta-review (59% for indeterminate FNAs and 86% for cytologically malignant FNAs). Because contraindications for thyroid lobectomy or near-total thyroidectomy are uncommon (e.g., pregnancy, inoperable tumor, or medically unable to undergo general anesthesia), the likely explanation for these lower-than-expected resection rates are that patients were lost to follow-up; that is, they were resected at an institution different from the study site performing the FNA. This can occur as almost all of the clinical sites are endocrinology and not surgical practices, and therefore it requires more effort from each clinical site to account for the patients' surgical disposition. Because we expected a cytologically malignant case to almost always undergo surgery, we estimated that 15% of cases were resected elsewhere (81% resection rate for Cyto M nodules in meta-review plus 15% equals a 96% "expected" resection rate). Adding this 15% factor to the 59% resection rate on indeterminate nodules, we estimated that only 74% of these patients get operated upon, and that the balance of patients and/or their physicians decided against surgery. The implication is that an estimated 25% of patients with indeterminate nodules are not undergoing surgical resection, and therefore some patients with cancer remain untreated. There is no evidence that unoperated patients have a lower risk of malignancy than operated patients in our study, as most physician investigators indicated at periodic telephonic follow-up that they intended for their patients with indeterminate or ma-

lignant cytology to be operated upon. In another investigation of patients with indeterminate nodules who did not undergo resection ($n = 637$) versus those that did ($n = 639$), no differences were found in age, sex, or race. Further, no differences were observed in the frequency of the most common FNA diagnosis (follicular neoplasm) or in the second most frequent diagnosis (suspicious for PTC), suggesting that there were no obvious clinical factors related to the decision not to undergo resection (26).

Local to expert panel pathologists comparison

In this series, surgical pathology slides for a subset of cases (221 cases) were centrally reviewed by a panel of two anatomical pathologist experts. The reviewed cases consisted of specimens from the entire FNA collection (i.e., prospectively collected in the clinic, pre- and postoperative specimens, and banked specimens) as the sample size of the operated prospective cohort alone for this analysis was relatively limited and we wanted to evaluate as many different neoplasm types as possible. Additionally, we found this larger set of cases to be comparable to the meta-review with respect to histology malignancy rate (46% and 40%, respectively), and therefore deemed this subset as reasonable to utilize for the surgical pathology concordance analysis.

Consistent with the published literature (37–39), there was relatively high interobserver variability between the local pathologists and the expert panel. When local histology was compared to the expert panelists, there was a benign to malignant categorical diagnostic disagreement of 10% between local pathologist and expert1 (observed agreement 90%, kappa = 0.79, CI 0.69–0.86), 13% between local pathologist and expert2 (observed agreement 87%, kappa = 0.75, CI 0.65–0.82), and 11% between local pathologist and expert panelist consensus (observed agreement 89%, kappa = 0.78, CI 0.69–0.86). This contrasted with lower disagreement rates of 8% between expert1 and expert2 preconferral (observed agreement 92%, kappa = 0.84, CI 0.77–0.90). The expert panelists disagreed on 17 cases preconferral; in 15 of 17 cases, the disagreement was follicular in nature (i.e., at least one of the two experts made the diagnosis of follicular or Hürthle cell adenoma, follicular or Hürthle cell carcinoma, or follicular variant papillary carcinoma). The kappa value for the comparison between expert panelists (0.84) was much higher than the comparison of local to expert panelists (0.75 and 0.79), meaning the expert panelists agreed with each other significantly more often than they agreed with the local pathologist. After expert1 and expert2 were unblinded to each other's diagnoses, they conferred and reached a consensus diagnosis on 97% of the cases, suggesting that a high degree of consensus is possible between experts on thyroid histopathological specimens. The rate of benign versus malignant

disagreement between local pathology and the expert panelists' consensus diagnosis was 11%, which underscores that central review by a panel of expert surgical pathologists should be utilized in studies evaluating accuracy of FNA test performance.

Because the expert panelists were able to achieve a high percentage categorical agreement with each other postconferral, their histopathological consensus diagnosis was regarded as a gold standard for comparison with local histopathology. The discordance between local and expert consensus diagnoses highlights the necessity for incorporation of an expert panel for central pathology review in clinical trial design for the development of thyroid diagnostic tests. The 11% disagreement rate between the local surgical pathology and expert panelists' consensus diagnoses suggests that a future molecular diagnostic test, which is developed using gold-standard histopathology, will likely be similarly discordant with local surgical pathology results.

Conclusion

The prospective subcohort in this study is the largest prospective, multicenter evaluation of thyroid FNA pathology to date. Postoperative risk of malignancy by cytopathological diagnosis for benign, indeterminate, and malignant thyroid FNAs was comparable to an updated meta-review of 11 large U.S.-based studies published from 2002 through 2010. Strengths of the current prospective specimen collection study include (i) a relatively homogeneous study population (98% U.S.-based, with age and gender similar across study sites), (ii) FNAs performed 99% with ultrasound guidance, and (iii) utilization of a surgical pathology diagnosis made by a panel of two external experts with high inter-rater diagnostic agreement both blinded and following conferral on discordant cases. Limitations of the current study include a small sample size on the FNA specimens that were both cytology benign and postoperatively malignant, and pending expert panel histopathology results for some cases with local surgical pathology. Limitations of the updated meta-review include variable mixes of UGFNA and PGFNA in each of the 11 published studies reviewed, as well as not having central expert pathologists re-evaluate and provide quality control for the surgical histology diagnoses.

FP results remain a concern for FNAs with indeterminate thyroid cytopathology, as the majority of these patients undergo surgery with 66% of the cases deemed histologically benign. The risk of malignancy in both the meta-review and the current prospective study was almost identical. In addition, both the prospective cohort and the meta-review found that approximately one quarter of patients with indeterminate nodules in the prospective clinically collected cohort appear to be opting out of thyroid resection, leading to a lack of appropriate surgery in a subset of patients with cancer.

In spite of recent guidelines seeking to standardize FNAB and interpretation of cytology results, FP and false-negative results continue to present a challenge in the evaluation of thyroid nodules. Molecular testing studies are needed to more accurately refine FNA diagnosis in the cytologically indeterminate group where the majority of cases prove to be benign and surgery could be avoided. Future molecular diagnostics studies should incorporate central review by experts in thyroid surgical pathology in their study design given the high variability in histopathological diagnosis with local pathologists.

References

1. Mazzaferri EL 1993 Management of a solitary thyroid nodule. N Engl J Med **328:**553–559.
2. Ross DS 2006 Editorial: predicting thyroid malignancy. J Clin Endocrinol Metab **91:**4253–4255.
3. Chen AY, Jemal A, Ward EM 2009 Increasing incidence of differentiated thyroid cancer in the United States, 1988–2005. Cancer **115:**3801–3807.
4. Seer Cancer Statistics Review 1975–2007. Available at http://seer.cancer.gov/csr/1975_2007/browse_csr.php?section=26&page=sect_26_table.05.html#a. Accessed July 16, 2010.
5. Davies L, Welch HG 2006 Increasing incidence of thyroid cancer in the United States, 1973–2002. JAMA **295:**2164–2167.
6. Zhu C, Zheng T, Kilfoy BA, Han X, Ma S, Ba Y, Bai Y, Wang R, Zhu Y, Zhang Y 2009 A birth cohort analysis of the incidence of papillary thyroid cancer in the United States, 1973–2004. Thyroid **19:**1061–1066.
7. Enewold L, Zhu K, Ron E, Marrogi AJ, Stojadinovic A, Peoples GE, Devesa SS 2009 Rising thyroid cancer incidence in the United States by demographic and tumor characteristics, 1980–2005. Cancer Epidemiol Biomarkers Prev **18:**784–791.
8. Caruso DR, Mazzaferri EL 1991 Fine needle aspiration biopsy in the management of thyroid nodules. Endocrinologist **1:**194–202.
9. Yassa L, Cibas ES, Benson CB, Frates MC, Doubilet PM, Gawande AA, Moore FD Jr., Kim BW, Nose V, Marqusee E, Larsen PR, Alexander EK 2007 Long-term assessment of a multidisciplinary approach to thyroid nodule diagnostic evaluation. Cancer **111:**508–516.
10. Lewis CM, Chang K-P, Pitman M, Faquin WC, Randolph GW 2009 Thyroid fine-needle aspiration biopsy: variability in reporting. Thyroid **19:**717–722.
11. Goldstein RE, Netterville JL, Burkey B, Johnson JE 2002 Implications of follicular neoplasms, atypia, and lesions suspicious for malignancy diagnosed by fine-needle aspiration of thyroid nodules. Ann Surg **235:**656–664.
12. Baloch ZW, Fleisher S, LiVolsi VA, Gupta PK 2002 Diagnosis of "follicular neoplasm": a gray zone in thyroid fine-needle aspiration cytology. Diagn Cytopathol **26:**41–44.
13. Bryson PC, Shores CG, Hart C, Thorne L, Patel MR, Richey L, Farag A, Zanation AM 2008 Immunohistochemical distinction of follicular thyroid adenomas and follicular carcinomas. Arch Otolaryngol Head Neck Surg **134:**581–586.

14. Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, Mazzaferri EL, McIver B, Pacini F, Schlumberger M, Sherman SI, Steward DL, Tuttle, RM 2009 Revised American Thyroid Association Management Guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid **19:**1167–1214.

15. Cibas ES, Ali SZ 2009 The Bethesda system for reporting thyroid cytopathology. Thyroid **11:**1159–1165.

16. Chudova D, Wilde JI, Wang ET, Wang H, Rabbee N, Egidio CM, Reynolds J, Tom E, Pagan M, Rigl CT, Friedman L, Wang CC, Lanman RB, Zeiger M, Kebebew E, Rosai J, Fellegara G, LiVolsi VA, Kennedy GC 2010 Molecular classification of thyroid nodules using high-dimensionality genomic data. J Clin Endocrinol Metab **95:** 5296–5304.

17. Dellelis RA, Lloyd RV, Heitz PU, Eng C 2004 World Health Organization Classification of Tumors: Tumors of Endocrine Organs. IARC Press, Lyon.

18. Williams ED, Abrosimov A, Bogdanova T, Ito M, Rosai J, Sidorov Y, Thomas GA 2000 Guest editorial: two proposals regarding the terminology of thyroid tumors. Int J Surg Pathol **8:**181–183.

19. Fleiss JL, Levin BA, Paik MC 2003 Statistical Methods for Rates and Proportions, third edition. Wiley-Interscience.

20. Blansfield JA, Sack MJ, Kukora JS 2002 Recent experience with preoperative fine-needle aspiration biopsy of thyroid nodules in a community hospital. Arch Surg **137:**818–821.

21. Sclabas GM, Staerkel GA, Shapiro SE, Fornage BD, Sherman SI, Vassillopoulou-Sellin R, Lee JE, Evans DB 2003 Fine-needle aspiration of the thyroid and correlation with histopathology in a contemporary series of 240 patients. Am J Surg **182:**702–710.

22. Castro MR, Gharib H 2005 Continuing controversies in the management of thyroid nodules. Ann Intern Med **142:** 926–931.

23. Wu HH, Jones JN, Osman J 2006 Fine-needle aspiration cytology of the thyroid: ten years experience in a community teaching hospital. Diagn Cytopathol **34:**93–96.

24. Yang J, Schnadig V, Logrono R, Wasserman PG 2007 Fine-needle aspiration of thyroid nodules: a study of 4703 patients with histologic and clinical correlations. Cancer Cytopathol **111:**306–315.

25. Oertel YC, Miyahara-Felipe L, Mendoza MG, Yu K 2007 Value of repeated fine needle aspirations of the thyroid: an analysis of over ten thousand FNAs. Thyroid **17:**1061–1066.

26. Banks ND, Kowalski J, Tsai HL, Somervell H, Tufano R, Dackiw APB, Marohn MR, Clark DP, Umbricht CB, Zeiger MA 2008 A diagnostic predictor model for indeterminate or suspicious thyroid FNA samples. Thyroid **18:**933–941.

27. Nayar R, Ivanovic M 2009 The indeterminate thyroid fine-needle aspiration. Cancer Cytopathol **117:**195–202.

28. Theoharis CG, Schofield KM, Hammers L, Udelsman R, Chhieng DC 2009 The Bethesda fine-needle aspiration classification system: year 1 at an Academic Institution. Thyroid **19:**1215–1223.

29. Faquin WC, Baloch ZW 2010 Fine-needle aspiration of follicular patterned lesions of the thyroid: diagnosis, management, and follow-up according to National Cancer Institute (NCI) recommendations. Diagn Cytopathol **38:**731–739.

30. Perros P, Clarke Susan EM, Franldyn J, Gerrard G, Harrison B, Hickey J, Kendall-Taylor P, McNicol AM, Mallick UK, Prentice M, Thakker RV, Watkinson J, Weetman AP British Thyroid Association, Royal College of Physicians 2007 Fine needle aspiration cytology. In: Guidelines for the Management of Thyroid Cancer, second edition. Lavenham Press, Lavenham, pp 9–10.

31. Bartolazzi A, Orlandi F, Saggiorata E, Volante M, Arecco F, Rossetto R, Palestini N, Ghigo E, Papotti M, Bussolati G, Martegani MP, Pantellini F, Carpi A, Giovagnoli MR, Moti S, Toscano V, Sciacchitano S, Penneli GM, Mian C, Pelizzo MR, Rugge M, Troncone G, Palombini L, Chiapetta G, Botti G, Vecchione A, Belloco R 2008 Galectin-3-expression analysis in the surgical selection of follicular thyroid nodules with indeterminate fine-needle aspiration cytology: a prospective multicentre study. Lancet **9:**543–549.

32. Baloch ZW, LiVolsi VA, Asa SL, Rosai J, Merino MJ, Randolph G, Vielh P, DeMay RM, Sidawy MK, Frable WJ 2008 Diagnostic terminology and morphologic criteria for cytologic diagnosis of thyroid lesions: a synopsis of the National Cancer Institute Thyroid Fine-Needle Aspiration State of the Science Conference. Diagn Cytopathol **36:**425–437.

33. Yeh MW, Demircan O, Ituarte P, Clark OH 2004 False-negative fine-needle aspiration cytology results delay treatment and adversely affect outcome in patients with thyroid carcinoma. Thyroid **14:**207–215.

34. Berner A, Sigstad E, Pradhan M, Groholt KK, Davidson B 2006 Fine-needle aspiration cytology of the thyroid gland: comparative analysis of experience at three hospitals. Diagn Cytopathol **34:**97–100.

35. Elisei R, Molinaro E, Agate L, Bottici V, Masserini L, Ceccarelli C, Lippi F, Grasso L, Basolo F, Bevilacqua G, Miccoli P, Di Coscio G, Vitti P, Pacini F, Pinchera A 2010 Are the clinical and pathological features of differentiated thyroid carcinoma really changed over the last 35 years? Study on 4187 patients from a single Italian Institution to answer this question. J Clin Endocrinol Metab **95:**1516–1527.

36. Ito Y, Miyauchi A, Inoue H, Fukushima M, Kihara M, Higashiyama T, Tomoda C, Takmura Y, Kobayashi K, Miya A 2010 An observation trial for papillary thyroid microcarcinoma in Japanese patients. World J Surg **34:**28–35.

37. Hirokawa M, Carney JA, Goellner JR, DeLelli RA, Heffess CS, Katoh R, Tsujimoto M 2002 Observer variation of encapsulated follicular lesions of the thyroid gland. Am J Surg Pathol **26:**1508–1514.

38. Franc B 2003 Observer variation of lesions of the thyroid. Am J Surg Pathol **27:**1177–1179.

39. Lloyd RV, Erickson LA, Casey MB, Lam KY, Lohse CM, Asa SL, Chan JKC, DeLellis RA, Harach HR, Kakudo K, LiVolsi VA, Rosai J, Sebo TJ, Sobrinho-Simoes M, Wenig BM, Lae ME 2004 Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. Am J Surg Pathol **28:**1336–1340.

Address correspondence to:
Chung-Che Charles Wang, M.D., M.B.A.
Department of Medical and Clinical Affairs
Veracyte, Inc.
7000 Shoreline Court, Suite 250
South San Francisco, CA 94080

E-mail: c.charles.wang@gmail.com