Errors in Thyroid Gland Fine-Needle Aspiration

Stephen S. Raab, MD,[1] Colleen M. Vrbin,[1] Dana Marie Grzybicki, MD, PhD,[1] Daniel Sudilovsky, MD,[1] Ronald Balassanian, MD,[1] Richard J. Zarbo, MD, DMD,[2] and Frederick A. Meier, MD[2]

Key Words: Cytology; Thyroid gland; Patient safety; Quality improvement

DOI: 10.1309/7RQE37K6439T4PB4

Abstract

Scant published data exist on redesigning pathology practice based on error data. In this first step of an Agency for Healthcare Research and Quality patient safety project, we measured the performance metrics of thyroid gland fine-needle aspiration, performed root cause analysis to determine the causes of error, and proposed error-reduction initiatives to address specific errors. Eleven cytologists signed out 1,543 thyroid gland aspirates in 2 years, and surgical pathology follow-up was obtained in 364 patients. Of the 364 patients, 91 (25.0%) had a false-negative diagnosis and 36 (9.9%) a false-positive diagnosis. Root cause analysis showed that major sources of error were preanalytic (poor specimen quality) and analytic (interpretation of unsatisfactory specimens as nonneoplastic and lack of diagnostic category standardization). We currently are evaluating the effectiveness of error reduction initiatives that target preanalytic and analytic portions of the diagnostic pathway.

The use of thyroid gland fine-needle aspiration (FNA) biopsy led to major improvements in the care of patients with thyroid gland nodules,[1-7] although with the widespread adoption of FNA, new challenges arose. These challenges included the optimization and standardization of diagnostic criteria[8] and the development of patient management protocols[1] based on diagnostic schema. Much of the thyroid gland FNA literature has established diagnostic criteria, documented performance metrics (eg, sensitivity and specificity), and highlighted the diagnostic pitfalls.[9] Our literature has not rigorously evaluated means to improve performance metrics or even seriously challenged the notion that current thyroid gland FNA error proportions are irreducible.

This study focuses on the development of a quality improvement initiative based on measuring baseline thyroid gland FNA error proportions. Our definition of a *diagnostic error* is a diagnosis that does not accurately represent the patient's actual disease process.[10,11] The sensitivity and specificity of thyroid gland FNA reportedly range from 57% to 99% and 90% to 99%, respectively.[2,9,12-17] These performance metrics depend on a number of factors, including the diagnostic categorical schema, availability of immediate interpretation, and operator experience.

We previously measured baseline multi-institutional anatomic pathology error frequencies and showed that thyroid gland FNA has relatively high false-negative and -positive proportions, compared with other specimen types, based on the cytologic-histologic correlation error-detection method.[11] In a 1-year period at 1 institution participating in an Agency for Healthcare Research and Quality–funded patient safety project, 26 of 231 thyroid gland FNA diagnoses were discrepant with surgical follow-up diagnoses (11.3% error frequency).[11]

Patients with thyroid gland FNA diagnostic errors had a portion (or all) of their thyroid gland removed for a benign condition or had a delay in diagnosis (and possibly excess testing) for failure to diagnose.[11]

In the present study, we used FNA performance metrics to drive root cause analysis to effect change and process redesign. Even though we believed that our thyroid gland FNA performance metrics were within the acceptable published range, our goal was to decrease the number of errors as much as possible. We evaluated all processes that contributed to error, and aspects of our process redesign are discussed. We are evaluating the affects of our process redesign, and these data are forthcoming.

Materials and Methods

Background

As previously mentioned, we conceived of this project when we found an 11.3% error frequency for thyroid gland FNA at 1 institution participating in an ongoing Agency for Healthcare Research and Quality patient safety project.[11] This error frequency was based on the cytologic-histologic correlation review of "2-step" or greater cytologic-histologic correlation diagnostic discrepancies.[10,11] Additional evaluation through existing institutional quality assurance practices showed that 23.1% of 2-step correlation discrepancies were diagnostic overcalls and 11.5% were diagnostic undercalls.[11] These errors were secondary to sampling (error in not obtaining diagnostic material), interpretation, or combined sampling and interpretation. The goals of the present study were to better define the problem of diagnostic error in thyroid gland FNA, perform root cause analysis, and redesign processes based on the findings of the root cause analysis.

Institutional review board approval was granted before the performance of the study.

Case Selection

To better define thyroid gland FNA error, we performed more detailed analysis of thyroid gland FNA diagnoses and outcomes for a 2-year period. We chose this timeframe to accumulate a sufficient number of cases to determine performance metrics for individual cytologists. One hypothesis that we examined was that differential use of diagnostic categories affected error frequencies. A search was performed of the laboratory information system (LIS; CoPathPlus Anatomic Pathology, Cerner, Kansas City, MO) to retrieve all thyroid gland FNA specimens examined between January 1, 2003, and January 1, 2005. We deidentified this institution for the purposes of this study. There were 1,343 patients and 1,543 FNA specimens; 147 patients had 2 FNAs, 19 had 3 FNAs, and 5 had 4 FNAs.

Based on the examination of the original thyroid gland FNA reports, the diagnoses were reclassified independently by 2 study investigators (S.S.R. and C.M.V.) into the following categories: unsatisfactory, benign, atypical, follicular or Hürthle cell lesion or neoplasm, "suspicious" for malignancy, or malignant.[10] Differences were resolved by consensus. In some cases, the original diagnoses were descriptive, and the investigators jointly evaluated and reclassified these diagnoses. Excluding the unsatisfactory diagnosis, this classification scheme is hierarchically scaled based on probability of neoplasia. We classified the follicular or Hürthle cell lesions and neoplasms in 1 category because the cytologists used these categories similarly. Some cytologists never used the neoplasm category, whereas others used the neoplasm and lesion categories.

We searched the LIS to retrieve surgical pathology follow-up for each patient who had a thyroid gland FNA specimen. The LIS was searched from the day of the FNA to June 30, 2005, and surgical pathology follow-up was obtained for 364 patients (27.1%). The follow-up timeframe ranged from the same day to 14 months after the thyroid gland FNA specimen was obtained; the mean follow-up time was 2.0 months. The final surgical pathology diagnosis was coded as benign, atypical, follicular or Hürthle cell adenoma, or malignant. We did not formally study the reasons for a surgical procedure in patients who had lesions diagnosed as benign by FNA; this was beyond the scope of this study. Patients who had a surgical pathology diagnosis of microscopic papillary carcinoma were coded as having a benign diagnosis for analysis because these diagnoses were incidental to the process.

Statistical Analysis

We measured overall and individual cytologist diagnostic performance metrics. Eleven cytologists interpreted the FNA specimens for the study period. For the overall FNA performance metrics, we calculated the number and frequency of cytologic-histologic correlation discrepant case pairs, sensitivity, specificity, likelihood ratio (LR) for individual diagnostic categories, and receiver operating characteristic (ROC) curves. For individual cytologist FNA performance metrics, we calculated the LR for diagnostic categories and ROC curves. We calculated individual cytologist differences in the proportion of their use of specific diagnostic category by using a χ^2 test, used to test for differences in proportions.

We calculated the number of discrepant cytologic-histologic correlation case pairs by measuring 2-step or greater differences between the FNA and surgical pathology diagnoses.[11] Because the FNA diagnostic category of atypical has different connotations, we calculated the number of discrepant case pairs in 2 ways: (1) excluding the atypical category and (2) considering an FNA diagnosis of atypical and a surgical pathology diagnosis of follicular or Hürthle cell neoplasm or malignant as discrepant.[18,19]

DOI: 10.1309/7RQE37K6439T4PB4

We recognized that all methods used to calculate sensitivity and specificity were biased because our thyroid gland diagnostic category scheme was nonbinary.[18] We calculated sensitivity and specificity by considering the surgical pathology diagnosis as the "gold standard," and, therefore, we excluded all cases without surgical pathology follow-up.[11] By using this exclusion criterion, we focused on cytologic-histologic discrepancy errors.[11] For calculation of sensitivity, we included FNA diagnoses of follicular or Hürthle cell lesion or neoplasm, suspicious, and malignant and surgical pathology follow-up diagnoses of follicular or Hürthle cell neoplasm or malignant as true-positives; we included an FNA diagnosis of benign and surgical pathology follow-up diagnoses of follicular or Hürthle cell neoplasm or malignant as false-negatives. For calculation of specificity, we included only an FNA diagnosis of benign and a surgical pathology follow-up diagnosis of benign as a true-negative; we included FNA diagnoses of follicular or Hürthle cell lesion or neoplasm, suspicious, and malignant and a surgical pathology follow-up diagnosis of benign as false-positives. We excluded patients who had FNA or surgical diagnoses of atypical or unsatisfactory from these calculations.

Our goal was to determine whether FNA could accurately classify lesions into 1 of 2 categories: lesions that should be excised (including carcinomas and all neoplasms) and those that do not need to be excised (benign, nonneoplastic lesions). We assumed that surgical excision was necessary to separate benign neoplasms from carcinomas.[11,19] We recognized the necessity for nondefinitive categories, although we wanted to determine whether cytologists were using nondefinitive categories differently.

We used the LR to express the probability of neoplasm for each FNA diagnostic category.[20-23] For each FNA diagnostic category, the LR is the quotient of the proportion of patients with disease who have a particular FNA diagnosis to the proportion of patients without disease who have that particular FNA diagnosis. The LR is a ratio of 2 probabilities, the probability of a test result if the disease is present (true-positives) divided by the probability of the same result if the disease is not present (false-positives). The LR has clinical value because it may be used to assess the posttest probability of disease if the pretest probability of disease is known.

The LR may range from 0 to infinity. An LR of less than 1.0 lowers the post-FNA probability of disease from the pre-FNA probability of disease. An LR equal to 1.0 does not alter the post-FNA probability of disease from the pre-FNA probability of disease. An LR of greater than 1.0 raises the post-FNA probability of disease from the pre-FNA probability of disease.

For the calculation of the LR, the surgical pathology follow-up diagnoses were coded as neoplasm present, which included follicular or Hürthle cell adenoma and malignant, or no neoplasm present, which included benign and atypical. An LR was calculated for each diagnostic category by dividing the proportion of cases with that diagnosis and follow-up indicating the presence of a neoplasm by the proportion of cases with that diagnosis and follow-up indicating no presence of a neoplasm. We calculated an LR for the diagnostic categories for the group of cytologists and for cytologists who had more than 30 specimens with surgical pathology follow-up.

We used ROC curves to express the diagnostic accuracy for the group of cytologists and for individual cytologists.[19,22,23] The ROC curve is an extension of the LR and is a graph of sensitivity vs (1 – specificity) as the cutoff value for the diagnosis is altered. A curve closest to the upper left hand corner of the graph is optimal, whereas a 45-degree line corresponds to random guessing. The SE was calculated with a nonparametric assumption. We calculated overall accuracy by measuring the area under the ROC curve. Compared with sensitivity and specificity measures, ROC curves incorporate the atypical and other nondefinitive diagnostic categories in measuring the overall diagnostic accuracy.

Root Cause Analysis

Three pathologists (S.S.R., D.M.G., and D.S.) performed root cause analysis to determine causes of error. Specimen, provider, and system factors all contributed to diagnostic pathology error. We used a root cause analysis method based on a modification of the Eindhoven Classification Model for the Medical Event Reporting System for Transfusion Medicine.[24-27] This method focuses on 3 domains: technical (equipment, forms, and software), organizational (procedures, policies, and protocols), and human (knowledge-based, rule-based, and skill-based). **Table 1** shows a more detailed list of the classification model.[24-27] The 3 domains were useful for classifying contributing factors and organizing causes of error. The domains also allowed for error investigation to focus on system factors rather than entirely on human factors.

We first performed root cause analysis by examining the overall and individual performance metric data and looking for causes of error based on less-than-optimal performance. We coded errors using the Eindhoven Classification Model, and we created a table displaying major factors that contributed to error. We realized that in much of clinical medicine, the most effective method of performing root cause analysis is immediately after the error occurs. Our method was somewhat limited because in some cases, root cause analysis was performed following a lengthy interval after the error occurred.[28,29] A benefit of studying test performance data is that system issues may be better studied.

We then proceeded to evaluate the cause of error in individual cases. For each 2-step cytologic-histologic discrepancy, we constructed a causal tree that visually represented the factors, activities, and decisions possibly leading to the diagnostic error. We recognized that all possible sources of error were not identifiable at the time we performed root cause analysis.

DOI: 10.1309/7RQE37K6439T4PB4

Table 1
Classification of Root Causes[24-27]

Code	Category	Definition
Latent errors	—	Errors that result from underlying system failures
Technical*		
TEX	External	Failures beyond the control of the investigating organization
TD	Design	Inadequate design of equipment, software, or materials; can apply to the design of workspace software packages, forms, and labels
TC	Construction	Designs that were not constructed properly, eg, incorrect setup and installation of equipment in an inaccessible area
TM	Materials	Material defects found, eg, weld seams on blood bags, defects in label adhesive, or ink smears on preprinted labels or forms
Organizational		
OEX	External	Failures beyond the control and responsibility of the investigating organization
OP	Protocols/procedures	Quality and availability of protocols that are too complicated, inaccurate, unrealistic, absent, or poorly presented
OK	Transfer of knowledge	Failures resulting from inadequate measures taken to ensure that situational or site-specific knowledge or information is transferred to all new or inexperienced staff
OM	Management priorities	Internal management decisions in which safety is relegated to an inferior position when there are conflicting demands or objectives; this is a conflict between production needs and safety
OC	Culture	A collective approach, and its attendant modes, to safety and risk rather than the behavior of just 1 person; groups might establish their own modes of function as opposed to following prescribed methods
Active errors	—	Errors or failures that result from human behavior
HEX	External	Failures originating beyond the control and responsibility of the investigating organization
Knowledge-based behaviors		
HKK	—	The inability of a person to apply his or her existing knowledge to a novel situation
Rule-based behaviors		
HRQ	Qualifications	The incorrect fit between a person's qualification, training, or education and a particular task
HRC	Coordination	A lack of task coordination within a health care team in an organization
HRV	Verification	The incorrect or incomplete assessment of a situation, including related conditions of the patient and donor and materials to be used, before beginning the task
HRI	Intervention	Failures that result from faulty task planning and execution; selecting the wrong rule or protocol (planning) or executing the protocol incorrectly (execution)
HRM	Monitoring	Failures that result from monitoring of process or patient status
Skill-based behaviors		
HSS	Slip	Failures in the performance of highly developed skills
HST	Tripping	Failures in whole-body movement; these errors are often referred to as "slipping, tripping, or falling"
Other factors		
PRF	Patient-related factors	Failures related to patient or donor characteristics or actions that are beyond the control of the health professional team and influence treatment
Unclassifiable	—	Failures that cannot be classified in any of the current categories

* Physical items such as equipment, physical installations, software, materials, labels, and forms.

In the cytologic-histologic correlation process, cytologists generally review slides to assign the error as sampling or interpretation (or both). Although we reviewed slides in this study, our root cause analysis method focused on the overall process, and we wanted to determine all the sources of error rather than simply classifying error as a clinical procurement or an interpretation problem. In the "Results" section, we provide 2 representative causal trees that depict causes of error appearing repeatedly. Finally, we listed 2 interventions that would address several of the major sources of error.

Results

Table 2 shows the surgical pathology follow-up diagnoses for each FNA diagnostic category. The numbers of 2-step cytologic-histologic discrepant case pairs excluding and including the atypical diagnoses were 93 (25.5%) and 127 (34.9%), respectively. By including the atypical and unsatisfactory diagnoses, 91 patients (25.0%) had a false-negative diagnosis and 36 (9.9%) a false-positive diagnosis. Of the false-negative diagnoses, 40 had carcinomas (30 papillary, 6 follicular, 2 Hürthle cell, 1 medullary, and 1 anaplastic) and the remainder had adenomas. Two patients had an FNA diagnosis of papillary carcinoma and benign surgical follow-up. Approximately 5 patients per month had a thyroid gland lobectomy for a benign condition or a delay in diagnosis for a neoplastic condition incorrectly classified. The FNA sensitivity and specificity were 70.2% and 67.0%, respectively. Of the FNA specimens diagnosed as benign, 41.8% had a follow-up surgical pathology diagnosis of follicular or Hürthle cell neoplasm or malignant.

Table 3 shows the number of specimens and the percentage of these specimens with surgical pathology follow-up by cytologist and FNA diagnostic category. More than 10% of

DOI: 10.1309/7RQE37K6439T4PB4

❚Table 2❚
Surgical Pathology Follow-up Diagnoses for Specific FNA Diagnoses

	Surgical Pathology Diagnosis				
FNA Diagnosis	Benign	Atypical	Follicular or Hürthle Cell Neoplasm	Malignant	Total
Unsatisfactory	6	0	6	0	12
Benign	71	0	30	21	122
Atypical	39	1	15	19	74
Follicular or Hürthle cell lesion or neoplasm	28	0	37	34	99
"Suspicious"	5	1	2	8	16
Malignant	2	0	0	39	41
Total	151	2	90	121	364

FNA, fine-needle aspiration.

❚Table 3❚
Fine-Needle Aspiration Diagnoses and Follow-up for Individual Cytologists[*]

Cytologist	Unsatisfactory	Benign	Atypical	Follicular Lesion or Neoplasm[†]	"Suspicious"	Malignant	Total
A	11 (0)	65 (2)	1 (0)	10 (80)	0	0	87 (10)
B	0	33 (12)	3 (67)	1 (100)	1 (100)	1 (100)	39 (23)
C	10 (10)	56 (13)	4 (50)	16 (94)	1 (100)	1 (0)	88 (30)
D	5 (20)	127 (16)	29 (72)	13 (77)	2 (100)	3 (67)	179 (31)
E	17 (18)	267 (10)	27 (52)	30 (60)	8 (75)	16 (88)	365 (22)
F	30 (17)	140 (13)	6 (50)	24 (79)	2 (100)	6 (100)	208 (25)
G	1 (0)	20 (10)	0	4 (100)	0	1 (100)	26 (27)
H	4 (0)	113 (12)	8 (63)	8 (88)	2 (100)	1 (100)	136 (21)
I	8 (25)	219 (7)	34 (50)	19 (63)	2 (50)	11 (91)	293 (20)
J	3 (0)	88 (15)	12 (75)	8 (63)	1 (100)	5 (100)	117 (28)
K	0	2 (100)	2 (50)	0	0	1 (100)	5 (80)
Total	89 (13)	1,130 (11)	126 (59)	133 (74)	19 (84)	46 (89)	1,543 (24)

[*] Data are given as number of cases with a specific fine-needle aspiration diagnosis (percentage of cases with surgical pathology follow-up).
[†] Also includes Hürthle cell lesions and neoplasms.

all patients who had a benign FNA diagnosis underwent a surgical procedure. A contingency table showed that the frequency of use of specific diagnostic categories depended on the individual cytologist ($P < .001$). For example, compared with the other cytologists, cytologist D used the atypical category more frequently and cytologist F used the unsatisfactory category more frequently. For cytologists D and F, the atypical category comprised 16% and 2.9%, respectively, of all diagnoses.

❚Table 4❚ shows that the LR for each FNA diagnostic category by cytologist had a wide range of variability. For example, the benign category for each cytologist had an LR that lowered the post-FNA probability of neoplasia, but to a considerably different percentage. The LR for the diagnostic category of atypical by cytologist usually lowered the post-FNA probability of disease, and for some cytologists, the LR for an atypical diagnosis was lower than the LR for a benign diagnosis.

❚Figure 1❚ shows the ROC curve for thyroid gland FNA for all cytologists combined. ❚Table 5❚ shows the area under the curve (diagnostic accuracy) for individual cytologists. Diagnostic accuracy varied considerably by cytologist. Excluding

❚Table 4❚
Likelihood Ratio for Fine-Needle Aspiration Diagnostic Categories by Cytologist

Cytologist	Unsatisfactory	Benign	Atypical	Follicular Lesion or Neoplasm[*]	"Suspicious"	Malignant	No. of Cases
All	0.73	0.52	0.62	1.84	1.21	14.1	364
A, B, C, G, H, K	∞	0.81	1.00	2.95	3.00	∞	83
D	0	0.67	0.75	4.00	1.0	∞	56
E	0.34	0.50	0.69	2.41	0.69	4.13	81
F	0.98	0.53	0.33	1.13	∞	∞	53
I	∞	0.30	0.58	1.97	∞	∞	58
J	—[†]	0.63	0.59	2.95	0	∞	33

[*] Also includes Hürthle cell lesions and neoplasms.
[†] Did not use the category.

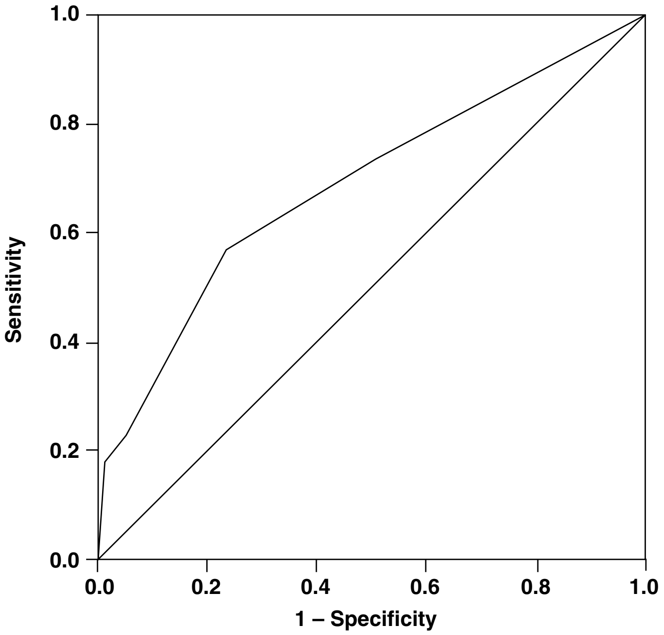

Figure 1 Receiver operating characteristic curve for all cytologists combined for thyroid gland fine-needle aspiration. The diagonal line represents chance alone.

Table 5
Area Under the Receiver Operator Characteristic Curve for Individual Cytologists

Cytologist	Area	SE	Asymptotic 95% Confidence Interval
A	0.667	0.218	0.240-1.093
B	0.583	0.198	0.196-0.971
C	0.643	0.115	0.418-0.867
D	0.661	0.073	0.518-0.804
E	0.685	0.060	0.568-0.803
F	0.670	0.073	0.526-0.813
G	0.583	0.248	0.096-1.070
H	0.700	0.102	0.500-0.900
I	0.743	0.064	0.617-0.869
J	0.673	0.093	0.790-0.856
K	1.0	0	1.0-1.0
Total	0.686	0.028	0.632-0.740

pathologist K, who saw few cases, only 2 pathologists (H and I) exhibited a higher diagnostic accuracy than the mean for the group.

Table 6 shows the results of the root cause analysis for the less-than-optimal performance characteristics. **Figure 2** and **Figure 3** show causal trees depicting sources of error. We identified sources of error in all domains, and diagnostic errors had a latent and an active component. The 2 main causes of error were false-negative diagnoses involving the interpretation of poor samples as nonneoplastic and false-positive diagnoses involving the interpretation of poor samples as neoplastic or nondefinitive. The majority of all errors involved sampling *and* interpretive components.

Based on these data, we chose 2 error reduction interventions: (1) creation of a standardized diagnostic terminology scheme that included standardized criteria to classify a specimen as adequate and (2) increasing the use of immediate interpretation FNA services. The first intervention involved

Table 6
Root Cause Analysis

Code*	Error Description
TD	Failure in laboratory equipment (eg, failure of centrifuge to properly prepare specimen); failure of radiology equipment during procedure (eg, ultrasound machine)
TC	Inappropriate setup of radiology suite, cytology laboratory, or FNA clinic to process specimens properly
TM	Defects in cassettes used for cell blocks; defects in needles, radiology equipment
OEX	Inability to reorganize cytology, radiology, and surgery services
OP	Lack of standardization for pathology sign-out procedures, diagnostic criteria for category use, radiology procedures
OK	New cytologists or less experienced cytologists not taught in a rigorous manner
OM	Radiology processes patients too quickly to allow for proper FNA performance; cytology schedule too busy to employ cytologists to perform immediate interpretation services; hospital does not mandate that patients with palpable lesions be sent to more experienced aspirators
OC	System focused on punishment and no improvement; no system for formal root cause analysis
HEX	Uncertain why specific patients with benign lesions were treated with surgical excision
HKK	Misdiagnosis of a case because of case-specific differences not previously observed
HRQ	Not all individuals have sufficient skills to perform FNA, interpret FNA, or manage patients with FNA diagnoses
HRC	Lack of coordination of radiology, pathology, and clinician teams
HRV	Incomplete assessment of situation (eg, aspiration performed of nonpalpable lesion in clinician office)
HRI	Incorrect specimen collection (eg, no smears prepared)
HRM	Immediate interpretation not performed
HSS	Poor performance of FNA or radiology equipment
HST	Incorrect material preparation (eg, material wiped off slide before immediate interpretation)
PRF	Disease not separable by FNA; patient desires benign lesions removed for reasons other than cancer risk; patient has propensity to yield poor specimen (ie, bleeding)

FNA, fine-needle aspiration.
[*] For an explanation of the codes, see Table 1.

an education component to improve diagnostic accuracy. Both interventions targeted the intersection of the preanalytic and analytic processes. These interventions are currently being tested.

Discussion

By using the cytologic-histologic correlation method of error detection, we showed that thyroid gland FNA at 1 institution had a false-negative proportion of 25.0% and a false-positive proportion of 9.9%. The root cause of most error was poor specimen quality. The patient had a neoplastic process and the FNA specimen was interpreted as adequate and nonneoplastic or the patient had a nonneoplastic process and the FNA specimen was interpreted as nondefinitive or neoplastic. In both scenarios, the lack of diagnostic standardization created "noise" that resulted in clinical confusion and overtreatment or undertreatment. The majority of errors were not caused by cytologist misinterpretation of good samples.

These data showed that the errors occurring in the diagnosis and care of patients with thyroid gland disease were multifactorial and interdisciplinary, crossing over radiologic (errors in tissue procurement), pathologic (errors in immediate and final interpretation and procurement), and clinical (interpretation of pathology results) services.[29] Error generally originated during specimen procurement, and cytologists perpetuated this error by interpreting poor specimens.

We recognized that our definition of error was rigid (a patient did not have neoplasia and the lesion was removed or the

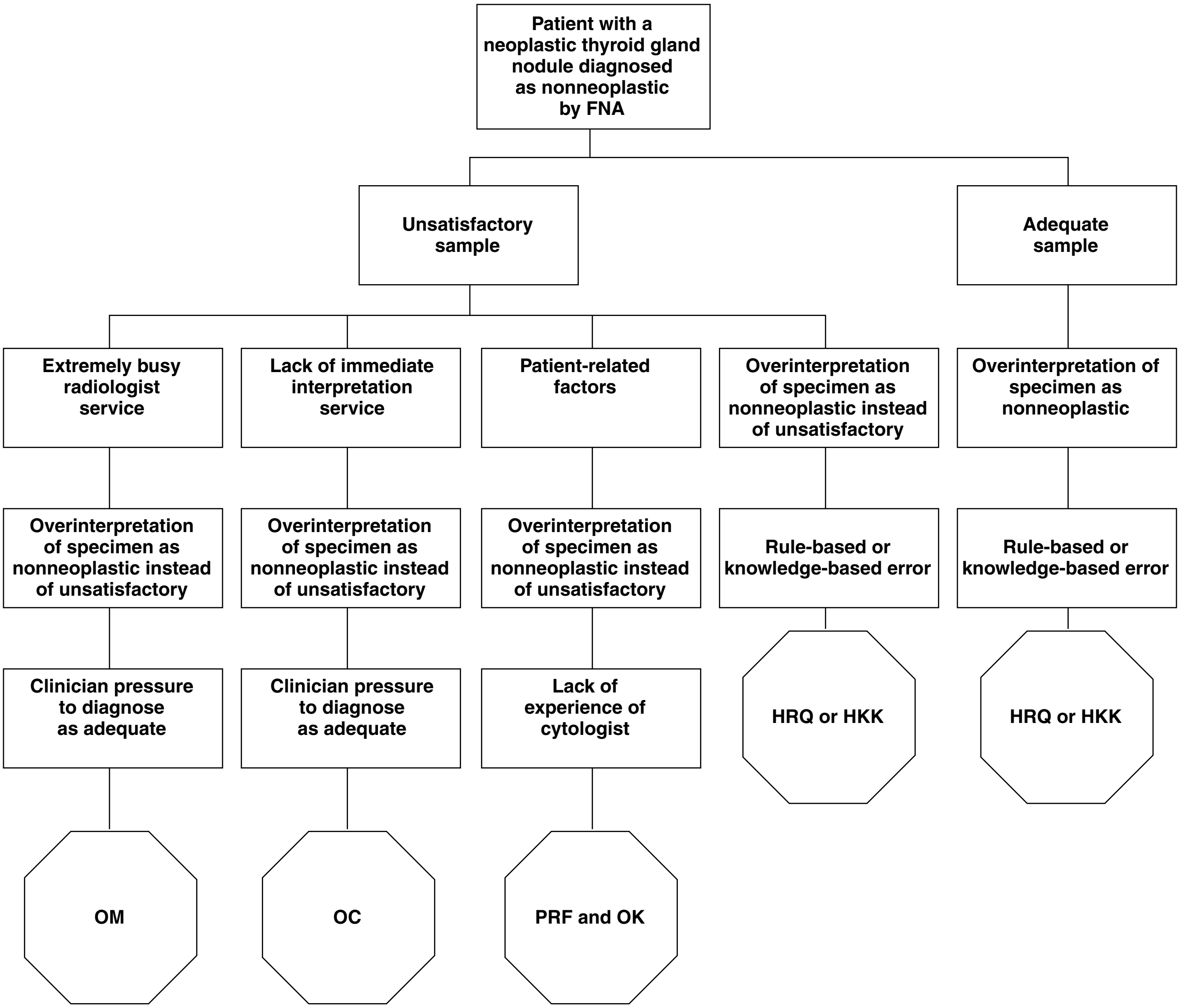
Patient with a neoplastic thyroid gland nodule diagnosed as nonneoplastic by FNA
Unsatisfactory sample
Adequate sample
Extremely busy radiologist service
Lack of immediate interpretation service
Patient-related factors
Overinterpretation of specimen as nonneoplastic instead of unsatisfactory
Overinterpretation of specimen as nonneoplastic
Overinterpretation of specimen as nonneoplastic instead of unsatisfactory
Overinterpretation of specimen as nonneoplastic instead of unsatisfactory
Overinterpretation of specimen as nonneoplastic instead of unsatisfactory
Rule-based or knowledge-based error
Rule-based or knowledge-based error
Clinician pressure to diagnose as adequate
Clinician pressure to diagnose as adequate
Lack of experience of cytologist
HRQ or HKK
HRQ or HKK
OM
OC
PRF and OK

Figure 2 The causal tree describing patients with a neoplastic thyroid gland nodule that is diagnosed as nonneoplastic by fine-needle aspiration (FNA). For an explanation of the codes in the bottom row, see Table 1.

DOI: 10.1309/7RQE37K6439T4PB4

patient had neoplasia and there was a delay in diagnosis).[10,11] Traditional diagnostic schema accept error as an unavoidable aspect of thyroid gland FNA; for example, the literature reports that up to 20% of patients with a diagnosis of a follicular neoplasm will have a nonneoplastic diagnosis on surgical pathology follow-up and that neoplastic lesions such as macrofollicular adenomas will exhibit "nonneoplastic" FNA features.[9] We accepted that errors secondary to patient-related factors may be irreducible (ie, some neoplastic and nonneoplastic conditions cannot be separated on FNA),[9] but we wanted to determine exactly what percentage of errors were not in this category.

Cytologists overinterpret unsatisfactory specimens as nonneoplastic for several reasons. First, cytologists use different criteria to diagnose an FNA specimen as satisfactory.[8,9,12,30-35] This practice is perpetuated by the lack of expert agreement in the cytology literature; some authors recommend more of some characteristics (eg, cells or colloid) than other authors. Clearly, the "minimum" necessary criteria are a sliding cutoff, and, if one accepts fewer of some criteria, one risks having a higher false-negative proportion.[9] However, the minimum necessary criteria debate centers on the belief that less-than-optimal specimens should be interpreted in the first place. If the entire diagnostic process were redesigned properly, cytologists would not even interpret such specimens. The root cause of this problem is mainly preanalytic but is augmented by cytologists who accept that they can interpret poor specimens. In the Toyota Production System model,[29,36] these specimens

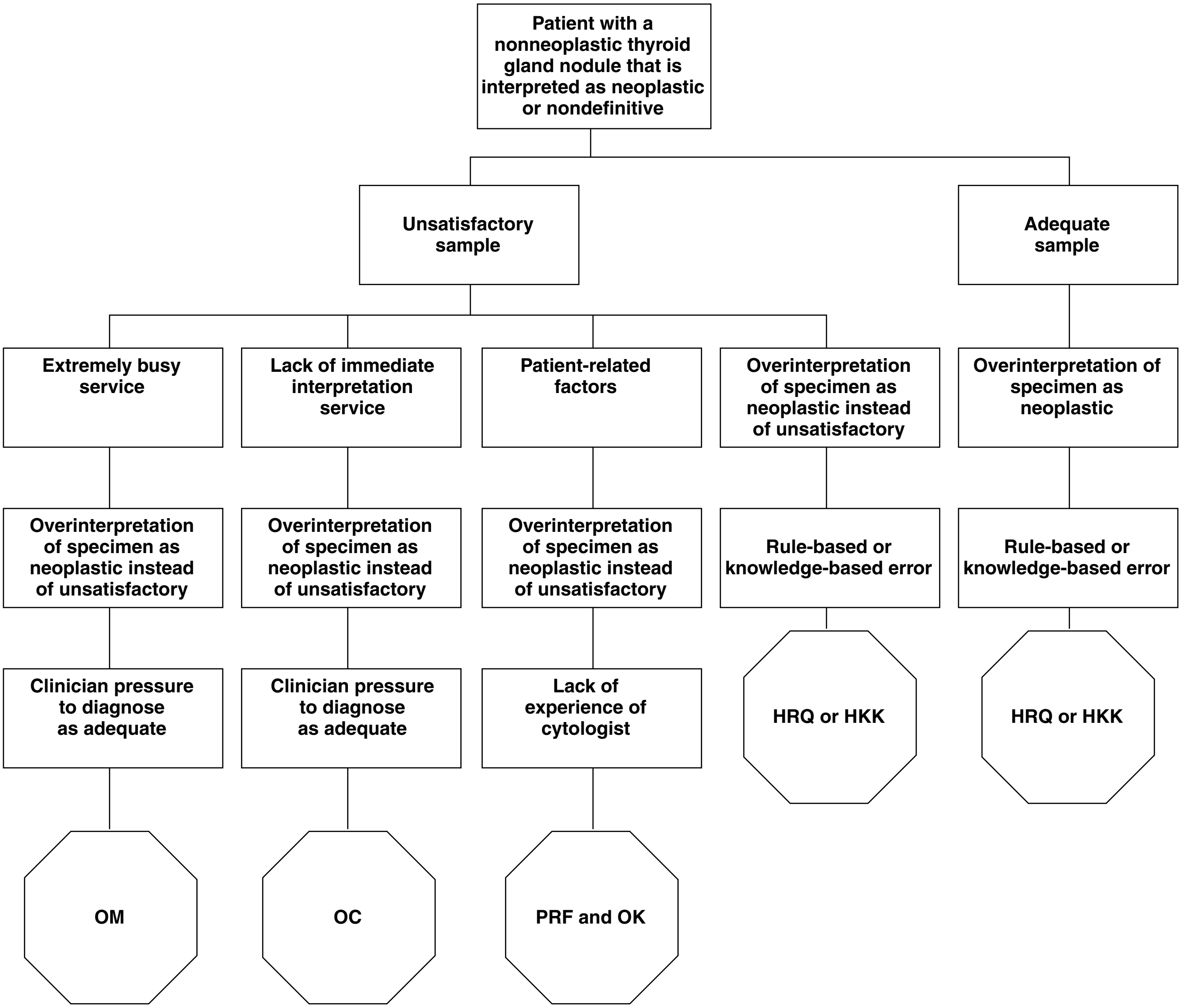
Patient with a nonneoplastic thyroid gland nodule that is interpreted as neoplastic or nondefinitive
Unsatisfactory sample
Adequate sample
Extremely busy service
Lack of immediate interpretation service
Patient-related factors
Overinterpretation of specimen as neoplastic instead of unsatisfactory
Overinterpretation of specimen as neoplastic
Overinterpretation of specimen as neoplastic instead of unsatisfactory
Overinterpretation of specimen as neoplastic instead of unsatisfactory
Overinterpretation of specimen as neoplastic instead of unsatisfactory
Rule-based or knowledge-based error
Rule-based or knowledge-based error
Clinician pressure to diagnose as adequate
Clinician pressure to diagnose as adequate
Lack of experience of cytologist
HRQ or HKK
HRQ or HKK
OM
OC
PRF and OK

Figure 3 The causal tree describing patients with a nonneoplastic thyroid gland nodule that is diagnosed as neoplastic by fine-needle aspiration. For an explanation of the codes in the bottom row, see Table 1

DOI: 10.1309/7RQE37K6439T4PB4

would be rejected as faulty. Probabilistically, cytologists are likely to be correct in interpreting these specimens as nonneoplastic, but in some cases (as shown herein), they will be incorrect. Underlying reasons for pathologists overinterpreting unsatisfactory specimens as nonneoplastic include clinical pressure and failures of knowledge-based or rule-based behaviors (eg, overconfidence).[37]

At this institution, cytologists used different criteria not only for the unsatisfactory category but also for other diagnostic categories. Root cause analysis showed that the cytologists used the diagnostic categories of atypical and follicular lesion or neoplasm differently. Some cytologists used the atypical category to classify lesions that exhibited some but not all features of papillary carcinoma (eg, abundant nuclear grooves). Other cytologists used this category to classify lesions that lacked specific criteria for a nonneoplastic or a neoplastic diagnosis but did not necessarily indicate that a papillary carcinoma was present.[9] The main underlying reasons for classification within this category were poor sampling (again a preanalytic error) and patient-related factors (eg, the lesion could not be accurately classified but represented a good aspirate). Because cases with this diagnosis were not separated into definitive neoplastic or nonneoplastic categories and because the cytologists wanted to convey different meanings, the clinicians behaved differently when provided an atypical FNA diagnosis. In practice, this diagnostic category lowered the post-FNA probability of neoplasia but created clinician confusion.[38] Root cause analysis also showed that clinicians contributed to the problem by pressuring some cytologists to use this category in a specified way.

We recognize that the performance metrics of thyroid gland FNA were biased because we chose to selectively examine cases in which surgical pathology follow-up was obtained. We did not obtain clinical follow-up for all patients with benign diagnoses. In this population, this bias resulted in a relatively high false-negative frequency. Our purpose in calculating accuracy in this manner was to highlight and study error rather than establish the performance metrics of FNA based on evaluation of clinical follow-up.

This study demonstrates that the cytologic-histologic correlation process was used to identify discrepancies in thyroid gland FNA to direct process change. Vrbin et al[39] and Raab et al[11] reported that the cytologic-histologic correlation practice has been performed variably across institutions, indicating that comparing discrepancy proportions across sites lacks a degree of validity. Raab et al[11] recommended standardization of the cytologic-histologic correlation process to effectively use it as a benchmarking tool. However, our study indicated that root cause analysis of cytologic-histologic correlation data helped identify repetitive causes of preanalytic and analytic error.

These errors pointed to a number of possible interventions, mainly directed at reducing the variability of specimen quality. First, the majority of the errors occurred in poor quality specimens procured by radiologists or clinicians. Increasing radiologist and clinician skill level is a difficult task for FNAs performed outside pathology, particularly if they are performed without immediate interpretation. For palpable lesions, several authors[9] have shown that pathologists have lower unsatisfactory proportions, although some clinicians are biased against switching to pathologist-driven FNA services. In addition, some pathology groups lack the desire or ability to perform thyroid gland FNAs and some clinicians are very skilled at obtaining excellent specimens. One of our interventions was increasing the number of pathologist-performed FNAs and immediate interpretation of clinician and radiologist-performed FNAs.

By developing more standard criteria for diagnosis, we targeted the preanalytic problem of specimen sampling and the analytic problem of diagnostic variability leading to clinician confusion. We targeted the sampling problem by adopting the strict use of specimen adequacy criteria that are based on optimally adequate specimens rather than minimally adequate specimens. We are measuring the success of these interventions, and our findings will be reported. Specifically, we will address how measures of specimen adequacy previously have focused on minimal numbers of cells and colloid and how arbitrary cutoff values of these numbers lead to errors.

From the Departments of Pathology, [1]University of Pittsburgh School of Medicine, Pittsburgh, PA, and [2]Henry Ford Health System, Detroit, MI.

Supported by grant HS13321-01 from the Agency for Healthcare Research and Quality, Rockville, MD.

Address reprint requests to Dr Raab: Dept of Pathology, University of Pittsburgh School of Medicine, 5230 Centre Ave, Pittsburgh, PA 15232.

References

1. Miller JM, Hamburger JI, Kini S. Diagnosis of thyroid nodules: use of fine-needle aspiration and needle biopsy. *JAMA.* 1979;241:481-484.
2. Goellner JR, Gharib H, Grant CS, et al. Fine needle aspiration cytology of the thyroid, 1980 to 1986. *Acta Cytol.* 1987;31:587-590.
3. Gharib H, Goellner JR. Fine-needle aspiration biopsy of the thyroid: an appraisal. *Ann Intern Med.* 1993;118:282-289.
4. Belfiore A, LaRosa GL, La Porta GA, et al. Cancer risk in patients with cold thyroid nodules: relevance of iodine intake, sex, age, and multinodularity. *Am J Med.* 1992;93:363-369.
5. Frable W. The treatment of thyroid cancer: the role of fine-needle aspiration cytology. *Arch Otolaryngol Head Neck Surg.* 1986;112:1200-1203.
6. Clark K, Moffat F, Ketcham A. Nonoperative techniques for tissue diagnosis in the management of thyroid nodules and goiters. *Semin Surg Oncol.* 1991;7:76-80.
7. Hamburger JI. Diagnosis of thyroid nodules by fine needle biopsy: use and abuse. *J Clin Endocrinol Metab.* 1994;79:335-339.

DOI: 10.1309/7RQE37K6439T4PB4

8. Papanicolaou Society of Cytopathology Task Force on Standards of Practice. Guidelines of the Papanicolaou Society of Cytopathology for the examination of fine-needle aspiration specimens from thyroid nodules. *Diagn Cytopathol.* 1996;15:84-89.
9. Geisinger K, Stanley MW, Raab SS, et al. Thyroid gland fine needle aspiration. In: *Modern Cytopathology.* Philadelphia, PA: Churchill Livingstone; 2004:731-780.
10. Raab SS. Improving patient safety by examining pathology errors. *Clin Lab Med.* 2004;24:849-863.
11. Raab SS, Grzybicki DM, Janosky JE, et al. Clinical impact and frequency of anatomic pathology errors in cancer diagnosis. *Cancer.* 2005;104:2205-2213.
12. Gharib H, Goellner J, Johnson D. Fine-needle aspiration cytology of the thyroid: a 12-year experience with 11,000 biopsies. *Clin Lab Med.* 1993;13:699-709.
13. Caruso D, Wester S, Kisken W. Fine-needle aspiration biopsy of solitary thyroid nodules: effect on cost of management, frequency of thyroid surgery, and operative yield of thyroid malignancy. *Minn Med.* 1986;69:189-192.
14. LaRosa G, Belfiore A, Giuffrida D. Evaluation of the fine needle aspiration biopsy in the preoperative selection of cold thyroid nodules. *Cancer.* 1991;67:2137-2141.
15. Caraway N, Sneige N, Samaan N. Diagnostic pitfalls in thyroid fine-needle aspiration: a review of 394 cases. *Diagn Cytopathol.* 1993;9:345-350.
16. Akerman M, Tennval J, Bioklund A. Sensitivity and specificity of fine needle aspiration cytology in the diagnosis of tumors of the thyroid gland. *Acta Cytol.* 1985;29:850-855.
17. Boey J, Hsu C, Collins R. A prospective controlled study of fine-needle aspiration and Tru-cut needle biopsy of dominant thyroid nodules. *World J Surg.* 1984;8:458-465.
18. Raab SS, Veronezi-Gurwell A. Thyroid nodules in the elderly: clinical management and incidence of malignancy as determined by fine needle aspiration biopsy. *Oncol Rep.* 1995;2:1151-1155.
19. Clary KM, Condel JL, Liu Y, et al. Interobserver variability in the fine needle aspiration biopsy diagnosis of follicular lesions of the thyroid gland. *Acta Cytol.* 2005;49:378-382.
20. Radack KL, Rouan G, Hedges J. The likelihood ratio: an improved measure for reporting and evaluating diagnostic test results. *Arch Pathol Lab Med.* 1986;110:689-693.
21. Beck JR. Likelihood ratios: another enhancement of sensitivity and specificity. *Arch Pathol Lab Med.* 1986;110:685-686.
22. Raab SS, Thomas PA, Lenel JC, et al. Pathology and probability: likelihood ratios and receiver operating characteristic curves in the interpretation of bronchial brush specimens. *Am J Clin Pathol.* 1995;103:588-593.
23. Raab SS, Bottles K, Cohen MB. Technology assessment in anatomic pathology: an illustration of technology assessment techniques in fine needle aspiration biopsy. *Arch Pathol Lab Med.* 1994;18:1173-1180.
24. Aspden P, Corrigan J, Wolcott J, et al. *Patient Safety: Achieving a New Standard of Care.* Washington, DC: National Academies Press; 2003.
25. Kaplan HS, Callum JL, Rabin Fastman B, et al. The Medical Event Reporting System for Transfusion Medicine: will it help get the right blood to the right patient? *Transfus Med Rev.* 2002;16:86-102.
26. Simmons D. Sedation and patient safety. *Crit Care Nurs Clin N Am.* 2005;17:279-285.
27. Kaplan HS, Battles JB, Van der Schaaf TW, et al. Identification and classification of the causes of events in transfusion medicine. *Transfusion.* 1998;38:1071-1081.
28. Dhir R, Condel JL, Raab SS. Identification and correction of errors in the anatomic pathology gross room. *Pathol Case Rev.* 2005;10:79-82.
29. Condel JL, Jukic DM, Sharbaugh DT, et al. Histology errors: use of real-time root cause analysis to improve practice. *Pathol Case Rev.* 2005;10:82-87.
30. Nguyen G-K, Ginsberg J, Crockford P. Fine-needle aspiration biopsy cytology of the thyroid: its value and limitations in the diagnosis and management of solitary thyroid nodules. *Pathol Annu.* 1991;26:63-91.
31. Kini S. In: Kline TS, ed. *The Thyroid: Guides to Clinical Aspiration Biopsy.* New York, NY: Igaku Shoin; 1987.
32. Renshaw AA. Evidence-based criteria for adequacy in thyroid fine needle aspiration. *Am J Clin Pathol.* 2002;118:518-521.
33. Schmidt T, Riggs M. Significance of nondiagnostic fine-needle aspiration of the thyroid. *South Med J.* 1997;90:1183-1186.
34. Burch H, Burman K, Reed H, et al. Fine needle aspiration of thyroid nodules: determinants of insufficiency rate and malignancy yield at thyroidectomy. *Acta Cytol.* 1996;40:1176-1183.
35. Kulkarni H, Kamal M, Arjune D. Improvement of the Mair scoring system using structural equations modeling for classifying the diagnostic adequacy of cytology material from thyroid lesions. *Diagn Cytopathol.* 1999;21:387, 393.
36. Raab SS, Andrew-Jaja C, Condel JL, et al. Improving Papanicolaou test quality and reducing medical errors by using Toyota Production System methods. *Am J Obstet Gynecol.* 2006;194:57-64.
37. Reason J. Human error: models and management. *BMJ.* 2000;320:768-770.
38. Powsner SM, Costa J, Homer RJ. Clinicians are from Mars and pathologists are from Venus. *Arch Pathol Lab Med.* 2000;124:1043-1046.
39. Vrbin CM, Grzybicki DM, Zaleski MS, et al. Variability in cytologic-histologic correlation practices and implications on patient safety. *Arch Pathol Lab Med.* 2005;129:893-898.

DOI: 10.1309/7RQE37K6439T4PB4