Effect of the Bethesda system for reporting thyroid cytopathology on thyroidectomy rates and malignancy risk in cytologically indeterminate lesions

Jennifer L. Rabaglia, MD, Wareef Kabbani, MD, Lucy Wallace, MD, Shelby Holt, MD, Lori Watumull, MD, Jeffrey Pruitt, MD, William H. Snyder, MD, and Fiemu E. Nwariaku, MD, Dallas, Texas

Background. Cytologically indeterminate thyroid nodules represent a diagnostic and therapeutic challenge. In 2007, the National Cancer Institute recommended The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) as a means of improving the accuracy of thyroid cytopathology. Our objective was to determine the effect of TBSRTC on thyroidectomy rates and malignancy risk in cytologically indeterminate lesions.

Methods. We compared thyroidectomy rates and malignancy risk in patients with indeterminate thyroid cytopathology across 2 time periods, spanning January 2000 and November 2009: pre-TBSRTC (January 2000 to September 2003) and post-TBSRTC (June 2008 to November 2009). Statistical comparisons were performed using the Fisher’s exact test and chi-square analysis ($P = .05$ significant).

Results. We performed 938 fine-needle aspirations in the first period, 765 in the second. We identified 78 (8.3%) cytologically indeterminate lesions in the pre-TBSRTC group and 91 (11.9%) lesions in the post-TBSRTC group. We found no difference in thyroidectomy rates between the groups (37/78 [47%] pre-Bethesda versus 32/91 [35%] post-Bethesda; $P = .12$). However, the malignancy rate was significantly lower in the post-TBSRTC group (13/37 [35%] pre-Bethesda versus 4/32 [13%] post-Bethesda; $P = .02$).

Conclusion. Application of TBSRTC is associated with lower malignancy risk in indeterminate thyroid nodules, despite similar thyroidectomy rates. These findings imply that standardization of cytologic classification improves diagnostic accuracy. (Surgery 2010;148:1267-73.)

From the University of Texas Southwestern Medical Center, Dallas, Texas

At some point during adulthood, 5–7% of American women and 1–2% American men will present with a palpable thyroid nodule. However, only 1 in 20 such nodules actually, represent a malignancy. Fine-needle aspiration (FNA) biopsy is the most cost-effective and accurate method available for assessment of thyroid nodules, with a sensitivity and specificity around 94% and 98.5% respectively for the diagnosis of malignancy. For a majority of patients, this technique represents a safe and accurate system for the triage of nontoxic nodules, determining whether operative intervention is indicated; however, approximately 15–30% of thyroid FNA biopsies are classified as indeterminate. This group encompasses a wide spectrum of thyroid nodular biology and represents an important clinical challenge. In addition, there is substantial intra- and inter-observer (pathologist) variation in the cytologic characteristics that are classified as indeterminate. These inconsistencies have led to confusion and affect the diagnostic accuracy of FNA in this subset of lesions.

In an attempt to address this problem, the National Cancer Institute sponsored a Thyroid FNA State of the Science conference, held in October of 2007 in Bethesda, Maryland, during which a group of multidisciplinary experts attempted to standardize the morphologic criteria and diagnostic terminology for the reporting of thyroid
FNA cytology. These efforts led to a 6-tiered diagnostic classification system named “The Bethesda System for Reporting Thyroid Cytopathology” (TBSRTC; Cibas and Ali2) The system was designed with 2 distinct purposes in mind: (1) to facilitate effective interdisciplinary and inter-institutional communication by standardizing the language of thyroid cytology and (2) to facilitate the decision-making process for direct caregivers by providing a set of diagnostic categories, each of which carries an explicit malignancy risk. A major change in the new classification was the inclusion of an additional category designated atypia of undetermined significance (AUS)/folicular lesion of undetermined significance (FLUS). This category is reserved for lesions with cellular or architectural characteristics that are not completely consistent with benign behavior but lack the sufficient severity or number of cytologic abnormalities to be classified as suspicious for either neoplasm or malignancy.

In June 2008, our institution adopted TBSRTC as the uniform method for reporting thyroid FNA results. The objective of this study is to determine the effect of TBSRTC on both thyroidectomy rates and malignancy risk within cytologically indeterminate lesions. Our goal was to determine whether the implementation of standardized diagnostic criteria and reporting language had any effect on malignancy rates in resected thyroid specimens.

METHODS

We performed a retrospective review of FNA records for patients undergoing biopsy of thyroid nodules at a single, large academic medical center comprised of multiple affiliated hospitals, during the period between January 2000 and November 2009. These records were divided into 2 distinct time periods: the “pre-Bethesda” group (January 2000 to September 2003) containing FNAs performed before the adoption of TBSRTC, and the “post-Bethesda” group (June 2008 to November 2009), consisting of FNAs performed and interpreted after institution-wide adoption of the TBSRTC. Demographic data including age and gender were collected in addition to cytologic results and subsequent follow-up (re-biopsy or histologic results).

No standardization for cytologic diagnosis was employed before the Bethesda guidelines were developed (ie, a diagnostic category was not mandatory). Instead, pathologists used varying descriptive terminology to categorize lesions. Not surprisingly, the greatest degree of variability was apparent with respect to cytology that was not clearly benign, malignant, or suspicious (for neoplasm or malignancy). Diagnoses in this ambiguous group predominantly included cellular follicular lesion (CFL), atypical (cell morphology or architecture), and to a lesser extent, microfollicular lesion. For the purposes of this study, much careful consideration was given as to which cytologic diagnoses should comprise the pre-Bethesda indeterminate category, in effort to maximize the validity of the comparison between the pre- and post-Bethesda data. After discussion with our pathologists, the cytologic categories included in our pre-Bethesda indeterminate category were atypical (cellular or architectural atypia), CFL, and microfollicular lesion.

The TBSRTC was adopted at our institution in June of 2008, and the post-Bethesda group consists entirely of patients biopsied after its implementation. In contrast with the previous approach, TBSRTC has provided a standard set of diagnostic criteria and terminology for the categorization of thyroid FNA cytology. The system consists of 6 diagnostic categories including nondiagnostic (unsatisfactory), benign, AUS/FLUS, suspicious for follicular neoplasm (FN), suspicious for malignancy, and malignant. A designation of AUS was reserved for aspirates with characteristics that were neither convincingly benign nor convincing of a neoplastic or malignant process.

Availability of surgical histology was used as a surrogate for operative rates, and malignancy rates were calculated based on frequency of malignancy among those cases with available histology. Incidental papillary microcarcinomas noted on final histopathology were classified as benign lesions, because these microcarcinomas are widely accepted as clinically insignificant findings. Demographic characteristics were compared between the pre- and post-Bethesda data sets for both the entire indeterminate cohort as well as between the operative subgroups. Differences in demographic characteristics were calculated using t-test or chi-square/Fisher’s exact tests as appropriate. Differences in operative rates and malignancy risk between pre- and post-Bethesda groups were calculated using Chi Square/Fisher’s exact test, with $P < .05$ indicating significance.

RESULTS

We examined 1703 thyroid FNAs obtained from the same number of patients for this study. These reports were divided into 2 time periods: the “pre-Bethesda” group consisting of 938 biopsies over the 33-month period between January 2000 and September 2003; and the “post-Bethesda” group consisting of 765 biopsies during the 18-month period between June 2008 and November 2009.
Patient demographics were similar between both time periods, both within the indeterminate cohort as well as the within operative subgroups. The only exception was mean age within the indeterminate cohorts, which was slightly greater in the post-Bethesda group (47 vs 54 years; \( P = .002 \)). This is summarized in Table I.

Of the 938 FNAs in the pre-Bethesda group, 78 (8%) fell into the indeterminate category. Of these, 37 (47.4%) had final histopathology available (indicating they underwent surgery for definitive diagnosis). The final histopathology was benign in 24 of 37 cases (64.9%; Fig 1, A and B). Thirteen of 37 patients (35.1%) in the pre-Bethesda group had malignant histopathology (Fig 1, C).

Of the 765 FNAs in the post-Bethesda data set, 91 specimens (12%) met the criteria for a diagnosis of AUS. Thirty-two of these 91 (35%) had final histopathology available, including 28 benign lesions (88%) and 4 malignancies (13%). The distribution of diagnoses is summarized in Fig 2. We found no difference in thyroidectomy rates (as indicated by availability of final histopathology) between the pre- and post-Bethesda groups (37/78 [47%] vs 32/91 [35%], respectively; \( P = .12 \)) (Table II). However, malignancy rates differed dramatically with the application of TBSRTC (13/37 [35%] in the pre-Bethesda group versus 4/32 [13%] in the post-Bethesda group; \( P = .02 \)). These results represent an effect size of 0.26 and a odds ratio of 1.7.

**DISCUSSION**

This study demonstrates that application of the Bethesda System for Reporting Thyroid Cytology resulted in lesser rates of malignancy within thyroidectomy specimens obtained after an indeterminate cytology report, without altering thyroidectomy rates. These findings imply that standardization of cytologic classification improves diagnostic accuracy and may eventually decrease the rate of nontherapeutic thyroidectomy.

FNA is an essential tool in the evaluation of euthyroid patients with thyroid nodules. FNA decreases the rate of unnecessary operations in patients with benign nodules, and helps to triage patients with definitive malignancy to the appropriate definitive operative procedure.\(^1,3\) Nevertheless, a major limitation of thyroid cytopathology is the poor performance of the test with certain types of lesions, most notably in follicular/Hurthle cell neoplasms. This limitation has led to the frequent utilization of an ‘indeterminate’ category, in which the cytologic characteristics of a nodule do not allow a designation of either benign or malignant. Further more, historically pathologists have varied widely in their application of criteria and terminology for the classification and reporting of thyroid FNA results.\(^4,5\) This inconsistency resulted in substantial variation in malignancy rates within this ambiguous group of lesions that, in turn, has generated some confusion among surgeons and endocrinologists regarding the optimal management of patients with indeterminate cytology.

A primary aim of the Bethesda panel was to provide standardized diagnostic criteria and reporting terminology for thyroid cytology, such that each well-defined cytologic category would carry an explicit risk of malignancy. This approach provides a common language to facilitate effective communication between pathologists, endocrinologists, and surgeons both within and across institutions. This also allows for clear risk stratification to be used in guiding clinical decision making within each diagnostic category. Given the inherent heterogeneity of indeterminate category before implementation of these criteria, it was unclear what effect this new approach would have on this group of lesions.

Historically, indeterminate lesions have accounted for 3–30% of thyroid cytologic specimens\(^1,5,7\); however, the exact cytologic characteristics defining ‘indeterminate’ have been highly variable.\(^4,7,8\) Some institutions included CFLs or FN, whereas others only consider lesions with cytologic or architectural atypia. We reported previously an 11% incidence of indeterminate cytology within thyroid FNA specimens from our institution; however, this group

### Table I. Patient demographics

<table>
<thead>
<tr>
<th></th>
<th>Pre-Bethesda</th>
<th>Post-Bethesda</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indeterminate cohort</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>78</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>47</td>
<td>54</td>
<td>.002 (S)</td>
</tr>
<tr>
<td>Mean</td>
<td>6–90</td>
<td>14–88</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td>.60 (NS)</td>
</tr>
<tr>
<td>Female</td>
<td>83%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>17%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Mean nodule size</td>
<td>3.4</td>
<td>3.0</td>
<td>.11 (NS)</td>
</tr>
<tr>
<td>Operative cohort</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>37</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Mean age (yrs)</td>
<td>47</td>
<td>53</td>
<td>.07 (NS)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td>.76 (NS)</td>
</tr>
<tr>
<td>Female</td>
<td>87%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>13%</td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td>Mean nodule size</td>
<td>3.8</td>
<td>3.3</td>
<td>.26 (NS)</td>
</tr>
</tbody>
</table>
included the atypical lesion, benign follicular lesion, CFL, and FN categories. In the current study, we have excluded benign follicular lesions and FNs from the pre-Bethesda indeterminate category, but included FNA specimens reported as microfollicular lesion. This approach was chosen in an attempt to most closely approximate the types of lesions that would fall into the newer AUS category. Furthermore, we believe that the management of benign follicular lesions and FN is fairly standard across institutions; therefore, these diagnoses do not represent a major area of confusion. The incidence of indeterminate lesions within the pre-Bethesda group as defined here was 8%, falling well within the 3–30% range previously mentioned. Within the post-Bethesda group, the incidence of AUS was approximately 12%, which is also consistent with historic rates of indeterminate cytology, but slightly greater than the 7% rate recommended by the Bethesda panel. The overuse of the AUS category is likely a reflection of a lack of confidence with the new diagnostic criteria, and we suspect that the proportion of AUS diagnoses will decline over time as user confidence levels increase.

The marked decrease in malignancy rate within cytologically indeterminate lesions after implementation of the TBSRTC is an interesting and promising finding. This effect is clearly independent of patient age and gender as well as lesion size and, therefore, suggests that the more clearly established diagnostic criteria utilized in TBSRTC resulted in a reclassification of cytologically ambiguous lesions. We hypothesize that the basis for this finding is 2-fold: first, cytologic specimens with more concerning characteristics were upgraded into the suspicious for FN category, leaving those with fewer or less concerning atypical features to remain within the AUS group; and second, the addition of a new, catch-all category may lead cytologists to overcall lesions that would have been considered benign previously. We would expect the second phenomenon to decrease with time as cytologists gain experience with the new criteria.

Fig 1. Distribution of histologic diagnoses in the pre-Bethesda indeterminate category. (A), Malignancy rate. (B), Distribution of diagnoses—benign. (C), Distribution of diagnoses—malignant.
system, whereas the first phenomenon is based more solidly on refined cytologic criteria and would be expected to endure. At this early post-adoption time point, the overall effect of the Bethesda system cannot be fully substantiated, but the trend looks promising. The TBSRTC seems to have successfully streamlined the classification of cytologically indeterminate lesions by defining a truly low risk indeterminate group with an expectedly low malignancy rate (meeting the preestablished goal of 5–15%).

It is not surprising that operative rates remained somewhat consistent between groups. This observation may be a reflection of the relative novelty of the TBSRTC, having not allowed for enough time to sustain a paradigm shift based on its application. Before the adoption of this system, the recommended management for lesions with indeterminate cytology was re-biopsy followed by operative intervention for diagnosis (hemithyroidectomy) if re-biopsy remained indeterminate.

This recommendation was based undoubtedly on the substantial and highly variable rates of malignancy noted within this diagnostic category in the past. Currently, experience with the new system is limited, and we continue to recommend thyroidectomy for the diagnosis of indeterminate lesions, because the malignancy risk within the new AUS category is as yet unproven. Based on the promising initial findings here, however, we would expect to perform fewer diagnostic thyroidectomies within the AUS category, perhaps utilizing a more selective approach for these lesions based on the presence or absence of associated suspicious clinical characteristics. The redefinition resulting from the TBSRTC may allow for a paradigm shift in the clinical management of these lesions, away from diagnostic thyroidectomy and toward surveillance, thereby decreasing the rates of nontherapeutic thyroidectomy.

Limitations of this study include the single-institution design, relatively limited number of patients, and limited time frame. In addition, histology was available in <50% of patients in both the pre-Bethesda indeterminate and post-Bethesda AUS groups, because fewer than half of these patients underwent thyroidectomy for diagnosis. Thus, the calculated malignancy rates provide only an estimation of the true rate of malignancy within these categories. To this end, the results of this study may provide an excellent basis for further investigation regarding the effects TBSRTC implementation across multiple institutions, perhaps providing definitive answers to these important clinical questions. Our findings imply that standardization of criteria and language for cytologic classification of thyroid cytology improves the diagnostic accuracy of FNA for cytologically indeterminate lesions, which may lead to a decrease in the rate of nontherapeutic thyroidectomy for thyroid nodular disease. Although great strides have been made with the introduction of TBSRTC, its adoption is far from universal and a concerted effort should be made to promote the adoption of this system as the standard of care across institutions.
I believe that if the projected malignancy rates within the Bethesda age were the same as those in the post-Bethesda age?

Also, just a general comment that our pathologists have informed us that their labs will be assessed annually by CLEA in terms of how they adhere to this classification scheme. What your thoughts are about how that will impact all of our surgical practices based on your analysis?

Dr Jennifer L. Rabaglia (Dallas, TX): To answer the first question, our cytopathologists were the same in both the pre- and post-Bethesda groups. However, there was some discrepancy between pathologists in the period between June of 2008 and the beginning of 2009, where we had adopted the Bethesda criteria but all pathologists had not yet adopted it uniformly. They achieved uniform adoption after January of 2009.

Second, I do think that strict adherence to these criteria will actually help to facilitate communication both among caregivers within a single institution as well as caregivers across multiple institutions, because it will provide a standardized language and a way for us to communicate our results in a more efficient fashion.

I believe that if the projected malignancy rates within the atypical (atypia of undetermined significance) category are borne out, most practitioners will adhere to the recommendation that these lesions be re-biopsied, and this may allow us to avoid unnecessary diagnostic lobectomy for some of these lesions.

Dr Christopher R. McHenry (Cleveland, OH): I very much enjoyed your presentation. And my question, I think you addressed in your last comment. But I want to understand.

So in a patient that has a FNA biopsy with atypia of undetermined significance, do you recommend—is your recommendation operation in all patients? Or do you use any clinical factors or sonographic features to help decide which patients you are going to be operating on?

Dr Jennifer L. Rabaglia (Dallas, TX): The current recommendation from the Bethesda panel is actually to re-biopsy lesions containing AUS or FLUS, and then to proceed to operative intervention for those who have atypia on a second biopsy and follow those who do not. That is our practice. There are 3 surgeons at our institution. Two of us practice in that fashion and 1 person still proceeds directly to thyroidectomy.

Based on our preliminary data here, I would say it is likely that we will all adopt the re-biopsy policy.

Dr Ashok R. Shaha (New York, NY): A very interesting paper and nice study. A couple of points.

I think it’s almost 2 years since the Bethesda system was published. It is still not adopted in many major medical centers. The Quest Lab and other labs, they don’t report it. Memorial does not report it according to the classification. They are reporting description-wise. And then as a clinician, we will have to see where it (the lesion) fits. So, I think there needs to be more propaganda about this classification so that there is a uniformity of reporting, which may take quite some time. That is 1 point.

The second thing, there’s always a concern, when you get the final report, about the incidental micropapillary carcinoma, which goes into greater category as 20 positive patients or 50 positive patients. But those incidental microcarcinomas have very little to do with the statistics of the needle biopsy. I think that information never comes out in many papers. We need to be quite cognizant about that, that needle the biopsy was not supposed to make a diagnosis of incidental microcarcinoma.

And the last 1 or 2 other points, I think which are very important, pertain to your last slide, which said that either we repeat the needle biopsy or we will reduce incidence of thyroidectomy. The decision about which patient should go to the operating room or not is not based on needle biopsy alone, especially indeterminate needle biopsy. It is based on a variety of factors, clinical factors: age, gender, size, patient’s anxiety. Diagnostic thyroid lobectomy is a common operative procedure for patient’s anxiety. But again, I think it is interesting to use these criteria and follow results for quite some time.

Dr Jennifer L. Rabaglia (Dallas, TX): I agree with all of those comments, the first comment being that this (application of the Bethesda Criteria) is actually not standardized procedure across the United States at this point, with some large academic centers adopting the criteria earlier than many others and most community hospitals. I do think that we need to get the word out to standardize these reads.
And it is true, in answer to Dr McHenry's second question and response to the comment, that the decision to operate on these lesions is definitely not based on cytology alone. It is a combination of patient factors, the appearance of a nodule on ultrasound (any suspicious characteristics), and of course patient anxiety factors in as well.

Dr Ashok R. Shaha (New York, NY): It's also a little confusing, the system they used—starting from 1, 2, 3, 4, 5, 6. In their system, 5 is cancer and 6 is nondiagnostic. Normally, it should go from nothing to something, to something more. I have no idea why they decided to put 6 as a nondiagnostic and 5 as cancer, which becomes quite confusing.