An International Interobserver Variability Reporting of the Nuclear Scoring Criteria to Diagnose Noninvasive Follicular Thyroid Neoplasm with Papillary-like Nuclear features. A validation study.

Running Title: Validation of Nuclear Scoring for NIFTP

Author: Lester D. R. Thompson, M.D.,[1] David N. Poller, MB ChB, M.D., FRCPath,[2] Kennichi Kakudo, M.D., Ph.D.,[3] William E. Gooding,[4] Raoul Burchette,[5] Yuri Nikiforova, M.D., Ph.D.,[6] Raja R. Seethala, M.D.[6]

[1]Southern California Permanente Medical Group, Woodland Hills, California, USA;

[2]University of Portsmouth, Department of Pathology, Queen Alexandra Hospital, Cosham, Portsmouth, UK;

[3]Department of Pathology and Laboratory Medicine, Nara Hospital, Kindai University Faculty of Medicine, Ikoma-city, Japan.

[4]University of Pittsburgh Cancer Institute, Pittsburgh, PA, USA;

[5]Research and Evaluation, Pasadena, CA, USA;

[6]Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA;

Address for page proofs, correspondence and reprints:

Lester D. R. Thompson, M.D.
Southern California Permanente Medical Group
Woodland Hills Medical Center
Department of Pathology
5601 De Soto Avenue
Woodland Hills, CA 91365

Tel: 818-719-2613; Fax: 818-719-2309; E-mail: Lester.D.Thompson@kp.org

Raoul J. Burchette: Raoul.J.Burchette@kp.org

William E. Gooding, University of Pittsburgh Cancer Institute, Biostatistics Facility

Suite 325 Sterling Plaza, 201 North Craig Street, Pittsburgh PA 15213

412-383-1583 weg@pitt.edu

Total counts: 1) Text pages: 23; 2) Tables: 7 3) Figures: 6

ACKNOWLEDGEMENTS

With appreciation for all pathologists who participated in this study, including 21 participants from the Southern California Permanente Medical Group and the Naval Medical Center-San Diego, 30 participants from various universities and hospitals in Japan, and 25 participants from the United Kingdom and Ireland, 20 of whom are members of the United Kingdom Endocrine Pathology Society.

Manuscript type: Original Research Article

Compliance with Ethical Standards:

Funding: There is no funding for this study.

Conflict of interest: All authors declare that they have no conflicts of interest. The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of Southern California Permanente Medical Group.

This article does not contain any studies with human participants performed by any of the authors and thus informed consent was not required.

ABSTRACT

Aim: To assess interobserver variation in reporting nuclear features of encapsulated follicular variant of papillary thyroid carcinoma, newly reclassified as non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP), based on a proposed standardized scoring system.

Methods: An education module was individually reviewed as a pre-evaluation teaching guide of the specific features of classical papillary carcinoma, the specific inclusion and exclusion features for the diagnosis of NIFTP, and a catalogue of the standardized scoring system of the nuclear features of papillary carcinoma used to reach this diagnosis. Participants subsequently reviewed 30 cases of thyroid lesions previously scored by members of the Endocrine Pathology Society Working Group for the Re-evaluation of the Encapsulated Follicular Variant of Papillary Thyroid Carcinoma. There was one uninvolved reference image to demonstrate fixation, processing and cell size and one image from each case for scoring, with results recorded for each participant. The location of training (country and program), years as a practicing pathologist, and approximate number of thyroid gland surgical cases diagnosed per year were recorded. The degree of agreement between participants was assessed by kappa statistics, using the individual criteria and the average, composite scores of the Working Group as a point of comparison.

Results: Using the Nuclear Standardized Scoring System, the interobserver agreement for final diagnosis score was generally excellent: unweighted and weighted kappa values between individual observers ranging from 0.242 to 0.930 (average 0.626). There was significant agreement between observers in reaching an interpretation of the presence or absence of nuclear features to diagnose NIFTP (score 0-1 versus score of 2-3), with California pathologists, 0.63 (median 0.66, SD 0.15), Japanese pathologists, 0.64 (median 0.66, SD 0.16) and UK pathologists, 0.60 (median 0.57, SD 014) compared to the expert panel, 0.70 (median 0.73, SD 0.19).

Conclusions: Use of the nuclear scoring system to evaluate the nuclear features of papillary thyroid carcinoma as applied to reach the diagnosis of NIFTP shows a good to substantial interobserver agreement, suggesting consensus can be reached in diagnosing the nuclear features required for this newly reclassified tumor.

Key Words: Thyroid neoplasms; Thyroid cancer, papillary; Carcinoma, papillary follicular/pathology; consensus; observer variation; scoring system; humans

INTRODUCTION

Papillary thyroid carcinoma (PTC) is the most common malignant neoplasm of the thyroid gland worldwide, usually associated with an excellent survival. Recently, "*The Endocrine Pathology Society Conference for Re-examination of the Encapsulated Follicular Variant of Papillary Thyroid Cancer*" convened March 20-21, 2015 in Boston, MA, and based on extensive evaluation of cases, outcome data and the development of a set of specific inclusion and exclusion criteria, issued a new name for this entity: ***Noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP)***.[1] During discussion by the 24 member experts, the notion of criteria reproducibility, specifically as it relates to the nuclear features of PTC was a major concern. Previous studies have shown considerable interobserver and intraobserver variability among experts in the diagnosis of the cytomorphonuclear features of papillary carcinoma in follicular variant tumors,[2, 3] with a range from 17 to 100%. Only 1 of 15 cases achieved unanimous agreement in one such review,[2] while concordance was achieved in 39% in another.[3] Thus, there seems to be a need to validate the nuclear features of PTC, perhaps including a hierachy of importance, and a qualitative assessment of these features in order to be reproducible, and hence achieve a more precise and potentially accurate overall diagnosis. Towards this end, only the nuclear features of PTC were examined in this validation, while the other architectural and cellular findings were not included (these criteria include: invasion, follicular architecture, papillae, psammoma bodies, necrosis, mitotic activity, encapsulation, colloid tincture, fibrosis and multinucleated giant cells or crystalloids). Specifically, the tumors showed a follicular pattern without papillae identified, so that architectural findings could not bias interpretation of the cytomorphonuclear features. Further, no minimum quantitative area within a tumor that showed nuclear features of PTC was established. One author has used at least

3 high power fields showing nuclear features of PTC per 3 mm of tumor diameter to establish bone fide nuclear features,[4] but this quantitation has not been tested. The are many cytomorphonuclear features of papillary carcinoma (Table 1). In discussion during the aforementioned consensus conference, nuclear alterations seem to get the most weighting by pathologists as they evaluate a case, but varying thresholds by each individual in application to an individual case, results in significant diagnostic variation. A standardized group of 3 major categories were defined by the consensus conference expert panel (Table 2), creating a binary “present” or “absent” value, and an assigned score of 0 or 1 for each major category. Therefore, nuclear size and shape; nuclear membrane irregularities; and nuclear chromatin characteristics are the three nuclear categories, with one point assigned if interpreted to be present. Overall, if there is a score of 2 or 3, then the nuclear features of PTC are sufficiently well developed as to be diagnostic (in whatever category tumor applied).

To validate and test the Nuclear Standardized Scoring System for papillary carcinoma nuclear features, cases were evaluated by general practicing pathologists in the United States, Japan and the United Kingdom (UK). The scores were then compared to the consensus conference experts, using individual scores, average scores and aggregates scores to determine a kappa reproducibility score.

Materials and Methods

The same set of 30 images used to initially develop and then ultimately revalidate the Nuclear Standardized Scoring System for papillary carcinoma nuclear features by the 24 members of the expert panel was used to test the reproducibility by general practicing surgical pathologist (see eTable 5 of Nikiforov, et.[1]). One pathologist was excluded at his request at the time of publication, and one pathologist did not rescore the validated criteria, but the original recorded data was still employed in the kappa statistic evaluation.

The nuclear features of papillary carcinoma were grouped into 3 categories (Table 2):

(1) nuclear size and shape: nuclear enlargement, elongation, overlapping and crowding (Figure 1);

(2) nuclear membrane irregularities: irregular contours, nuclear grooves, nuclear folds, intranuclear cytoplasmic pseudoinclusions (Figure 2); and

(3) nuclear chromatin characteristics: chromatin clearing, margination to the membranes, glassy nuclei and fine even delicate chromatin (Figure 3).

A 3-point scoring scheme assigned each class of nuclear features a score of 0 or 1, with a range of scores from 0 to 3. Based on a mutation positive endpoint serving as the reference standard for the original test set (n=18 cases with molecular testing) [1], a score of 0 or 1 was considered inadequate for the diagnosis of NIFTP (i.e. benign), while a score of 2 or 3 was considered sufficient for the diagnosis of NIFTP (see supplemental table 1 for additional information). The same visual guide used by the expert panel (see eFigure 4 of Nikiforov, et.[1]) was included at the start of the evaluation.

An invitation to participate was extended to 30 general surgical pathologists in California (United States of America), Japan and the United Kingdom (UK) by the authors (LDRT, KK, DNP).

The instructions were as follows:

1) Study the teaching module, highlighting the inclusion and exclusion criteria for the diagnosis (about 45-60 minutes);

2) Attempt to review the 30 cases in one or two seatings (i.e., over a short interval, estimating 30 minutes to complete);
3) Enter a score of “0” or “1” for each category;
4) Provide demographic data, including years of experience and volume of work;
5) Aggregated and individual results will be compared to the 24 endocrine expert pathologists for statistical evaluation of reproducibility;

A 72-page training module (PDF) was distributed with the invitation (an expanded version can be viewed via: DOI: 10.13140/RG.2.1.2063.9124), along with a spreadsheet that included the case number and 3 columns in which to score the three nuclear feature categories (Table 2). The “total score” was automatically calculated using the standard Excel spreadsheet sum function. Additional recorded data included the number of whole years as a practicing pathologist (post training), residency training program attended, fellowship training and/or certification, board certification and/or additional board certification, and a reasonable approximation of total number of thyroid surgical pathology cases diagnosed/signed out per year.

Completed spreadsheets were received from 21 pathologists in the greater Southern California region, 30 pathologists in Japan, and 26 pathologists from the UK, which comprised the basis for all further analysis. Each groups’ demographics are summarized in Table 3. A few explanatory notes regarding the demographics are highlighted in supplemental Table 2 for sake of completeness.

Previously published molecular data served as the reference standard to fit a random-effects logistic regression model to predict molecular diagnosis based on molecular status and individual pathologist’s nuclear score as previously described [1]. The logistic model accounted for correlation among pathologists evaluating the same case. Based on prior validation {Nikiforov, 2016 #6064, score 0-1 was used to identify mutation-negative and score 2-3 mutation-positive lesion.

The received data were also analyzed using a kappa statistic (McNemar statistic) for all 30 cases for each of the 24 expert pathologists, and each of the individual pathologists from each of the 3 geographic regions, including calculation of a super kappa (weighted Kappa). Mean, median, range and standard deviation was calculated for each group. Kappa statistics were also calculated for each of the criteria for the experts and well as for each participant, with averages, medians and standard deviations. A super kappa was calculated as a sum of the numerators of the individual kappas divided by the sum of the denominators of the individual kappas for each reviewer as a form of weighted kappa. The results were based on comparing each reviewer to the expert consensus for each criteria and also for the overall diagnosis, while the experts were also compared to their own consensus. Standard commercial available software (statistical package for the social sciences, SPSS) was used for these calculations, using confidence intervals of 95% for all positive findings and an alpha level set at $p<0.05$.

The kappa statistic (kappa coefficent) uses 1 to indicate perfect agreement and 0 indicating agreement by pure chance alone. The following kappa statistics are used for interpretation in this paper:

Kappa	**Agreement**
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

RESULTS

Each pathologist from each geographic region provided a 0 or 1 value for each of the 3 nuclear categories. For example, a reference of normal, uninvolved adjacent thyroid gland parenchyma from the same slide captured at the same magnification was provided (Figures 4-6), while a single, representative high power field (20x magnification) using an Aperio scanned digital image from 30 different cases were included for review. These were the same cases reviewed by the expert panel, although the images included in this publication are unique from those already published, included for illustrative purposes only.

The overall average score for each case by geographic region (Table 4) was calculated, showing, on average, a substantially similar result for each case between the individual groups when compared to the experts. The UK pathologists scored cases with an overall lower average total score than the experts, Californian or Japanese pathologists. As each case was individual, a standardized mean could not be achieved, but suffice it to say there was an overall bias towards a lower total score among UK pathologists.

Across all pathologists, utilizing 0 or 1 to be diagnostic of a benign nodule and a score of 2 or 3 to be diagnostic of NIFTP, the nuclear scoring system demonstrated: a sensitivity of 75.0% (95% CI, 72.2%-77.6%), specificity of 76.9% (95% CI, 72.4%-80.8%), and overall accuracy of 75.5% (73.2%-77.7%). The accuracy of the model was not significantly influenced by country, years of experience or number cases seen. However, the UK pathologists trended towards lower accuracy, while there was a very slight trend towards improved accuracy with increased experience and number of cases (see supplemental Table 3, and supplemental Figures 1, and 2).

Overall, a kappa statistic was rendered for each of the individual criterion and as a predictor of final diagnosis (Table 5), with a super kappa calculated as a weighted average. Again, the UK pathologists tended to show a lower overall mean score for each criterion, with the first criterion (nuclear size and shape) the most likely to be difficult to achieve a meaningful kappa statistic. However, when all three criteria were aggregated to achieve a final diagnosis score (either 0/1 or 2/3 total score), then a substantial agreement was achieved, with a kappa statistic from 0.6 to 0.64 for the general surgical pathologists in comparison to 0.70 for the expert pathologists. The super kappa statistic yielded a kappa of 0.55 to 0.56 in comparison to 0.61 for the experts, again a finding that showed moderate to substantial agreement. Overall, the bin midpoint kappa values suggested a good to substantial agreement (Table 6) for each of the geographic areas, a finding that was similar to the results for the experts. When compared as overall kappa statistics between the regions, there was significant reproducibility (Figure 7).

DISCUSSION

This validation study has demonstrated a reasonable likelihood that general surgical pathologists will be able to apply the criteria of the NIFTP Nuclear Standardized Scoring System to an individual case and achieve a result that is in good to substantial agreement with expert pathologists. Reproducibility is important in the ability to apply specific criteria, and that was the main aim of this validation. Thus, the remaining criteria (invasion, patterns of growth, area of tumor with the nuclear features of PTC) were not evaluated in this cohort.

Accuracy of the previously established cut-off using the cases in this set with molecular results as a reference standard as previously calculated appears to be somewhat lower as compared to that of the original expert scoring {Nikiforov, 2016 #6064}. This is not an entirely unexpected, since in addition to "expertise," the expert group was additionally primed repeatedly over the course of several months by continued teleconferences and thus visual repetition of specific nuclear features that likely improved standardization of thresholds within the group. Even with a diagnostic accuracy of ~75% when extending to this international group of 77 pathologists, the performance of this scoring system is still improved from the historical precedent in prior (albeit smaller) studies of follicular patterned lesions [2]. Thus even with some deprecation from the accuracy of the original expert group, the performance of the nuclear scoring system is still favorable.

Neither, number of years in practice, number of cases reviewed, nor location significantly factored into the accuracy of the model. However, as expected, there is slight a trend towards a higher number of cases (experience) per year and a higher number of years in practice being correlated to being able to recognize the nuclear features with greater consistency and reproducibility, and thus "accuracy" in final diagnosis. However, the estimated number of cases is fraught with bias, and the exact number of years in practice versus years of specialization or years post training is different for each of the locales. Interestingly, there is a trend towards reduced scoring by the UK pathologists. It is well known that there is geographic variation in applying thresholds for follicular patterned lesions [3, 5]. UK and European pathologists have an overall different training and threshold for the interpretation of the nuclear features of papillary carcinoma, and thus this lower scoring is to be expected. However, when tasked with reviewing the education module before scoring, there may be a change or increase in scoring. This trend will be open to further discussion as greater experience with the diagnostic term and its application is developed.

It seems that with education and experience, "*The Endocrine Pathology Society Conference for Re-examination of the Encapsulated Follicular Variant of Papillary Thyroid Cancer*" criteria for scoring the nuclear features considered diagnostic of papillary-like nuclear features can be applied with reproducibility and yield similar diagnostic results between experts and general practicing pathologists for the newly reclassified ***Non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP)***.

CONCLUSION

There is good to substantial agreement between general surgical pathologists around the globe and expert endocrine pathologists in achieving reproducible Nuclear Standardized Scoring System for papillary carcinoma nuclear features total scores for the classification of the nuclear features of papillary carcinoma. This NIFTP Nuclear Standardized Scoring System can thus be applied with confidence in approaching thyroid gland nodules or tumors in attempting to yield a reproducible result between independent reviewers of these nuclear features.

References

1. Nikiforov YE, Seethala RR, Tallini G, et al. Nomenclature Revision for Encapsulated Follicular Variant of Papillary Thyroid Carcinoma: A Paradigm Shift to Reduce Overtreatment of Indolent Tumors. JAMA Oncol. 2016;2: 1023-1029.

2. Elsheikh TM, Asa SL, Chan JK, et al. Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. Am J Clin Pathol. 2008;130: 736-744.

3. Lloyd RV, Erickson LA, Casey MB, et al. Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. Am J Surg Pathol. 2004;28: 1336-1340.

4. Thompson LDR. Ninety-four cases of encapsulated follicular variant of papillary thyroid carcinoma: A name change to Noninvasive Follicular Thyroid Neoplasm with Papillary-like Nuclear Features would help prevent overtreatment. Mod Pathol. 2016;29: 698-707.

5. Hirokawa M, Carney JA, Goellner JR, et al. Observer variation of encapsulated follicular lesions of the thyroid gland. Am J Surg Pathol. 2002;26: 1508-1514.

TABLES

Table 1: Cytomorphonuclear Features of Papillary Thyroid Carcinoma.

Enlarged cells (compared to adjacent parenchyma)
Increased (high) nuclear to cytoplasmic ratio
Nuclear enlargement
Nuclear overlapping
Nuclear crowding (high cellularity)
Nuclear elongation
Irregular nuclear placement within the cell (luminal, mid or basal)
Nuclear grooves or folds
Nuclear contour irregularities (demi-lunes, rat-bites)
Intranuclear cytoplasmic psuedoinclusions
Pale, fine delicate nuclear chromatin distribution
Nuclear chromatin clearing (Orphan Annie nuclei)
Nuclear chromatin margination/condensation at the periphery
Prominent nuclear membranes
Multiple nucleoli
Nucleoli identified on the nuclear membrane

Table 2: Nuclear Features of Papillary Thyroid Carcinoma Scoring Criteria.*

Nuclear features (1 point each)	**Criteria**
1. Size and shape	Enlargement, elongation, crowding, overlapping
2. Membrane irregularities	Irregular contours, grooves, folds, intranuclear cytoplasmic inclusions
3. Chromatin characteristics	Chromatin clearing, margination to the nuclear membranes, glassy nuclei, fine-even delicate chromatin
Score:	0: absent or only slightly expressed 1: present or well developed
Total score:	0 or 1: Not diagnostic 2 or 3: Diagnostic of papillary thyroid carcinoma nuclei

*Modified from Nikiforov, et. al. [1]

Table 3: Demographics of General Surgical Pathologists

Characteristic	**United States of America**	**Japan**	**United Kingdom**
Total Number	21	30	26
Years of practice:			
Range	1 – 36	1 – 31	1 – 41
Mean	13.6	15.6	14.9
Median	10.0	14.5	15.0
Number of unique residency training programs	13	n/a	7
Fellowship or additional training	16	28	8
Cytology	5	28	6
Surgical Pathology	4	n/a	2
Other	7	n/a	n/a
Board certification	21	30	20
Estimated/actual annual number of thyroid cases			
Range	6 – 69	5 – 150	10 – 300
Mean	20.2	37.5	96.9
Median	14	30	67.5

Table 4: Overall Mean Score for Each Case by Geographic Group Compared to the Expert Panel

Case #	Experts	USA	Japan	UK
A8	**2.958**	3.000	2.800	2.731
A26	**2.500**	2.762	2.600	2.192
A27	**2.042**	1.952	2.000	1.462
A29	**1.125**	1.286	1.300	1.385
A35	**2.708**	2.476	2.633	2.423
A36	**2.375**	1.619	1.967	1.692
A37	**0.083**	0.000	0.133	0.115
A38	**1.625**	1.524	1.267	1.538
A41	**1.208**	1.000	0.800	0.962
A43	**1.125**	0.952	0.867	0.500
A46	**0.667**	0.476	0.400	0.077
A47	**0.792**	1.143	1.133	0.731
A52	**0.167**	0.048	0.100	0.154
A56	**1.042**	0.619	0.800	0.423
A58	**1.333**	1.238	1.367	1.077
A59	**2.167**	2.238	2.033	1.923
A60	**0.000**	0.333	0.167	0.115
A62	**2.708**	2.810	2.600	2.615
A73	**2.750**	2.810	2.467	2.577
A79	**2.292**	2.143	2.133	1.577
A80	**0.708**	0.952	0.667	0.423
A102	**3.000**	2.952	2.767	3.000
A111	**2.792**	3.000	2.733	2.846
A120	**1.750**	1.810	1.367	0.923
A121	**0.542**	0.429	0.600	0.269
A126	**2.792**	2.667	2.333	2.269
A127	**2.833**	2.381	2.400	2.385
A128	**2.583**	2.571	2.333	2.231
A134	**2.083**	2.095	2.067	1.962
A136	**1.917**	1.905	1.533	1.192

Table 5: Kappa Statistic for Individual Criteria and Final Diagnosis between Groups and the Experts

Kappa Statistic	Values	Experts	California	Japan	UK
Kappa for 1st Criterion	Median	0.60	0.47	0.48	0.44
	Mean	0.54	0.49	0.41	0.44
	St. Dev	0.19	0.15	0.20	0.14
Kappa for 2nd Criterion	Median	0.67	0.61	0.67	0.66
	Mean	0.61	0.60	0.63	0.65
	St. Dev	0.26	0.15	0.21	0.19
Kappa for 3rd Criterion	Median	0.67	0.61	0.60	0.59
	Mean	0.68	0.60	0.59	0.57
	St. Dev	0.14	0.16	0.2	0.16
Kappa for final diagnosis (score of 0/1 vs 2/3)	Median	0.73	0.66	0.66	0.57
	Mean	0.70	0.63	0.64	0.60
	St. Dev	0.19	0.15	0.16	0.14
Super Kappa	Median	0.64	0.60	0.57	0.55
	Mean	0.61	0.56	0.55	0.55
	St. Dev	0.13	0.09	0.12	0.11

Table 6: Kappa Value Bin Midpoints by Regional Pathologists

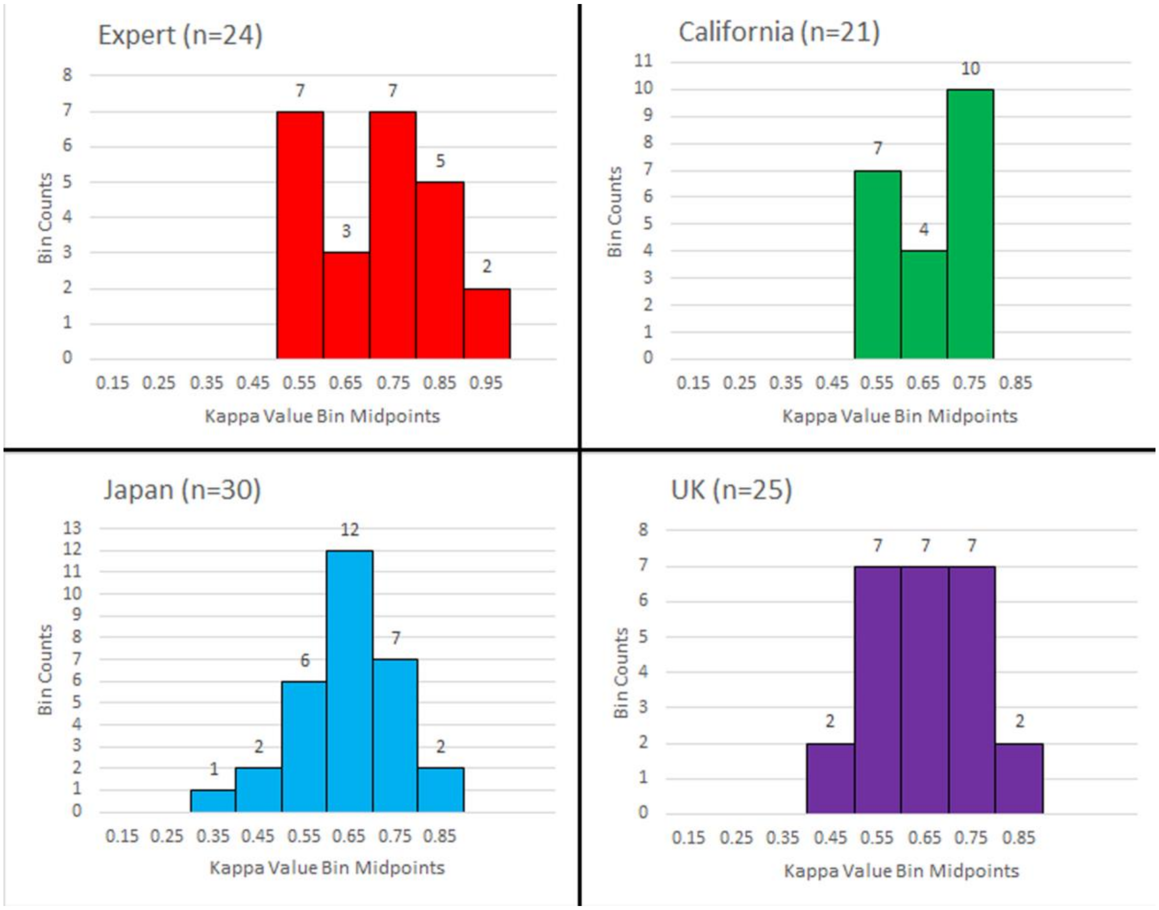


Table 7: All Kappa Values Counts by Regional Pathologists

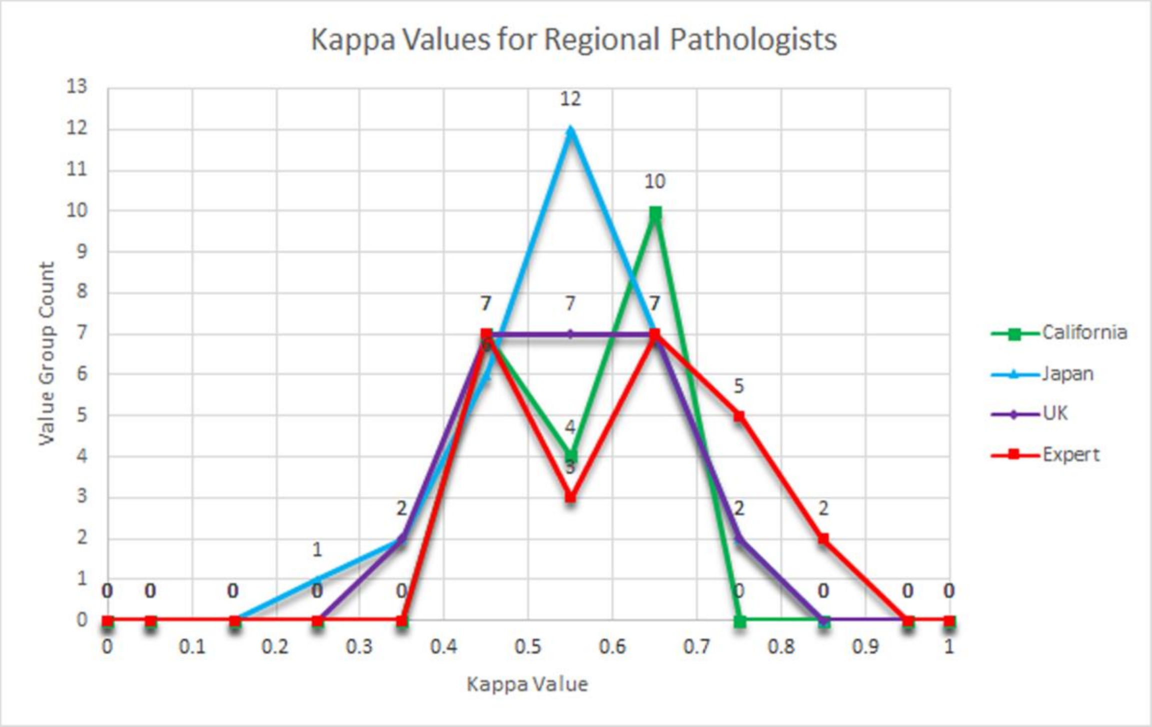


FIGURE LEGENDS

Figure 1: The cells are enlarged, with a very high nuclear to cytoplasmic ratio, showing greatly enlaraged nuclei, showing elongation and overlapping (case A102).

Figure 2: The nuclei show contour irregularities, nuclear folds and nuclear grooves (cases A102 and A008, left and right respectively).

Figure 3: Nearly all of the nuclei in this field show optical clearing, with margination of the nuclear chromatin towards the periphery (case A111).

Figure 4: Case A008. The inset shows the size of the normal cells and nuclei. This case had an overall average nuclear score of 2.844, well above the 2 cutoff for nuclear features of papillary carcinoma.

Figure 5: Case A037. The inset shows the size of the normal cells and nuclei. This case had an overall average nulcear score of 0.083, a total score interpreted to be benign.

Figure 6: Case A111. The inset shows the size of the normal cells and nuclei. This case had an overall average nuclear score of 2.86, a total score intepreted to represent papillary nuclear features.

SUPPLEMENTAL TABLE 1

For modeling purposes in the original evaluation, using a molecular end point as the reference "gold standard" separating NIFTP from benign conditions, the scoring scheme achieved the most accurate classification when a score of 0 or 1 was diagnostic of a benign nodule (mutation negative, n=5) and a score of 2 or 3 was diagnostic of NIFTP (mutation positive, n=13).

- Sensitivity - 86.5% (82.7% - 90.3%)
- Specificity - 80.8% (73.8% - 87.9%)
- PPV - 92.2% (89.1% - 95.2%)
- NPV - 69.8% (62.2% - 77.4%)
- Classification Accuracy - 85.0% (82.8% - 90.3%)

Note: Three of the expert pathologists had scores that were consistently higher with respect to the overall mean, and two pathologists had identical scores for all cases evaluated, a bias perhaps accounted for by joint review.

SUPPLEMENTAL TABLE 2

The participants were asked to estimate the number of surgical pathology thyroid cases they signed-out/rendered a diagnosis on each year. Respondents from Japan and the UK had their clinical practice in a community hospital, a regional medical center, a referral center or an academic center, while all of the California participants were in a community practice setting.

For California participants:

- The exact number of surgical pathology thyroid gland cases was retrieved from a centralized surgical pathology reporting database (CoPath)
- The average estimated number was 29.4, while the actual average number was 20.2 cases per pathologist per year indicative of a marked bias to over-estimation of the actual number of cases reviewed
- A similar extrapolation may apply to the other cohorts and serves as a limitation

For UK participants:

- The Royal College of Pathologists in Cellular Pathology (FRCPath) exam is equivalent to the American Board of Pathology Certification.
- Most of the respondents for the UK were either more interested in or experienced in endocrine organ pathology, as 20 or the 26 respondents were members of the UK Endocine Pathology Society. Pathologists are referred to as specialists when they have additional training, but official registration only started in 1996, and thus official records cannot substantiate personally provided data.
- True years of experience versus years of practice may be skewed for various reasons.

1. Initial responses from UK pathologists included training (cellular pathology training is equivalent to anatomic pathology residency in the USA and is about 5-6 years on

average), which was not done for the other groups. Thus, this figure had to be clarified and reduced accordingly.

2. Further, specialist registration was voluntary initially, and some pathologists never registered.
3. Foreign graduates are handled unique to the country where training was obtained: European Union countries have specialization data entered effective from the date of specialization in their home country, but consultants who immigrated from Asia or the Middle East have to apply to the UK General Medical Council with the date of specialist accredidation recorded only from the point of application forward.

For Japanese and UK participants

Actual case count may be skewed for several reasons.

1. Pathologists in Japan and the UK routinely show thyroid gland cases to other members of their department or associated hospitals, and as such, the actual number of specimens reviewed is different from cases signed out or co-signed, potentially accounting for the significantly higher case load for the UK pathologists than the other groups.
2. In addition, some pathologists review a higher number of thyroid gland cases as a result of second-read or external referral of the patient to a new hospital/region, and local custom dictates a review of all surgical pathology material before additional treatment is implemented locally.
3. Finally, there is an increasing trend towards subspecialist practice, especially in the UK, where a selected person(s) at the cancer center reviews all thyroid gland cases, not only for their own patients, but also for local-regional hospitals or cancer centers; thus, one

thyroid gland case may be reviewed by up to 5 pathologists, all of whom count it as a case diagnosed.

Regarding designation as an expert

- Expert designation is highly subjective
- However, definitive separation of an "expert pathologist" from a pathologist with significant years of experience or a high volume (>200 cases per year) practice cannot be reliably determined. Referral or second opinion consultation cases in endocrine organ pathology sent to a specific individual who has a significant publication and teaching record would qualify the person as an expert; these criteria were not definitively met by the participants in study.
- While not a significant factor, a few of the respondents (up to 5) may have published in endocrine organ pathology, although without a significant publishing track record (i.e., more than 2 publications per year in a peer reviewed Medline indexed journal) or major education role (lecturing or teaching on endocrine organ pathology for medical students, residents or practicing pathologists at regional, national or international venues).

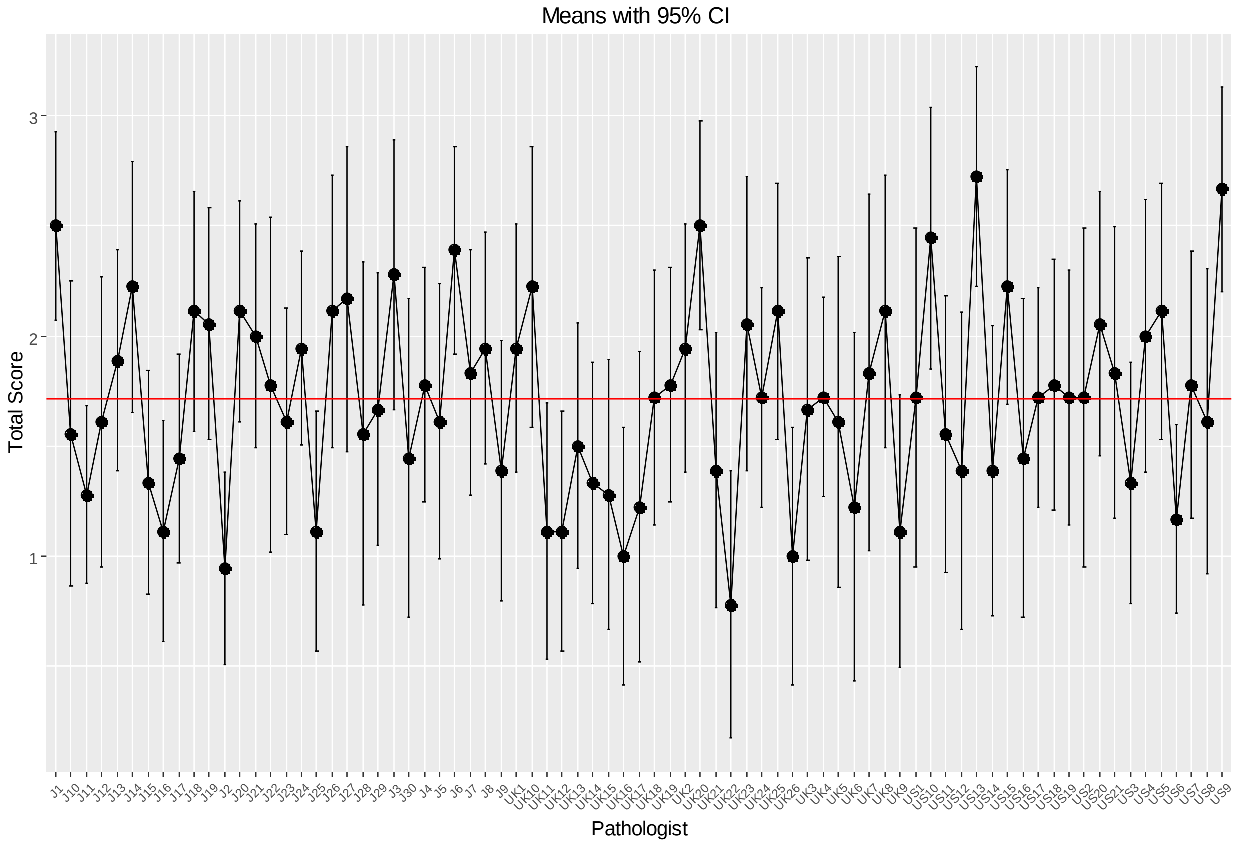
SUPPLEMENTAL TABLE 3

Validation of all 77 observers from each country using 0 or 1 to be diagnostic of a benign nodule and a score of 2 or 3 to be diagnostic of NIFTP was as follows:

- Sensitivity - 75.0% % (72.2%-77.6%)
- Specificity – 76.9% (72.4% - 80.8%)
- PPV – 89.4% (87.1%-91.3%)
- NPV – 54.2% (50.0%-58.3%)
- Classification Accuracy – 75.5% (73.2% - 77.7%)

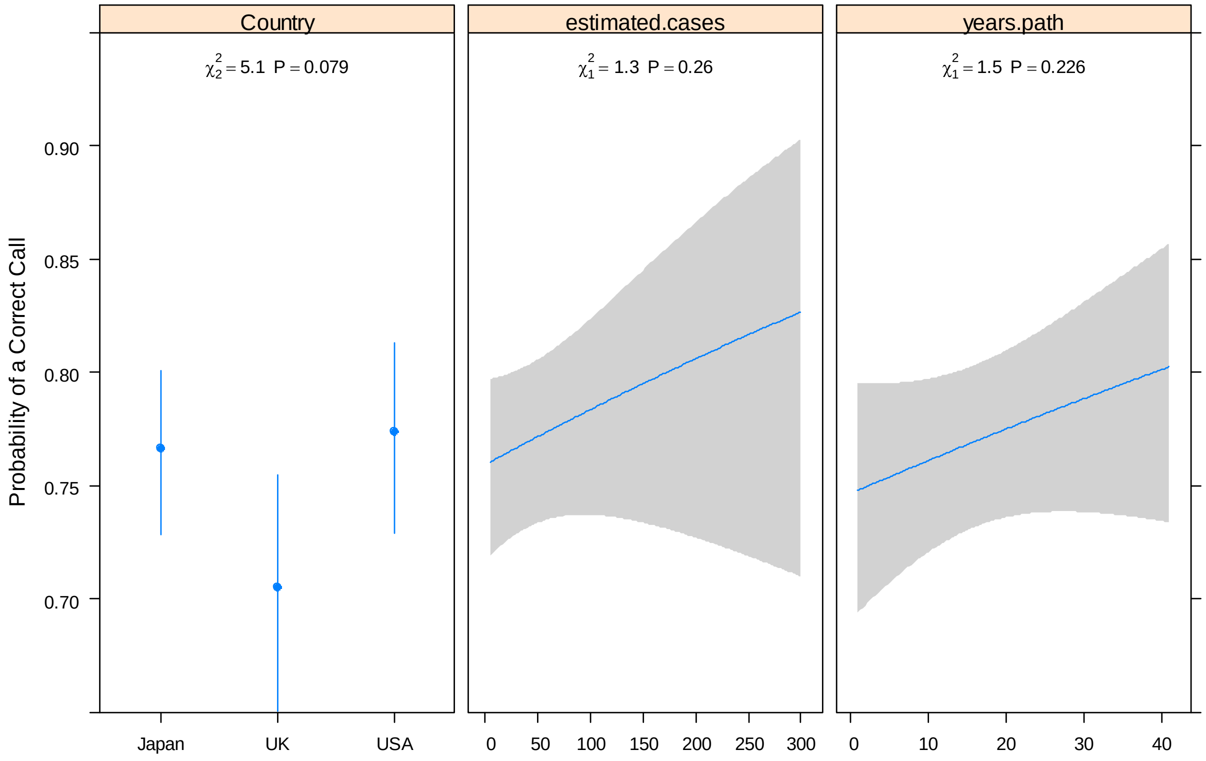
SUPPLEMENTAL FIGURE 1

Mean case score with 95% confidence interval by individual pathologist for cases in the training set. The red horizontal line is the grand mean across all cases and all pathologists. Many UK pathologists are well below the mean.



SUPPLEMENTAL FIGURE 2

Diagnostic accuracy as a function of country and experience. No parameter was significant, but UK pathologists show lower accuracy, and both number of cases and years in practice do appear to improve accuracy, albeit marginally.


Country
$\chi^2_2 = 5.1$ P = 0.079
estimated.cases
$\chi^2_1 = 1.3$ P = 0.26
years.path
$\chi^2_1 = 1.5$ P = 0.226
Probability of a Correct Call
Japan
UK
USA