JPTM
Journal of Pathology and Translational Medicine 2024; 58: 299-309
https://doi.org/10.4132/jptm.2024.07.25

ORIGINAL ARTICLE

Diagnostic challenges in the assessment of thyroid neoplasms using nuclear features and vascular and capsular invasion: a multi-center interobserver agreement study

Agnes Stephanie Harahap[1,2], Mutiah Mutmainnah[3], Maria Francisca Ham[1,2], Dina Khoirunnisa[4], Abdillah Hasbi Assadyk[5], Husni Cangara[6], Aswiyanti Asri[7], Diah Prabawati Retnani[8], Fairuz Quzwain[9], Hasrayati Agustina[10], Hermawan Istiadi[11], Indri Windarti[12], Krisna Murti[13], Muhammad Takbir[14], Ni Made Mahastuti[15], Nila Kurniasari[16], Nungki Anggorowati[17], Pamela Abineno[18], Yulita Pundewi Setyorini[19], Kennichi Kakudo[20]

[1]Department of Anatomical Pathology, Faculty of Medicine, Universitas Indonesia/Dr. Cipto Mangunkusumo Hospital, Jakarta;
[2]Human Cancer Research Center-Indonesian Medical Education and Research Institute, Faculty of Medicine, Universitas Indonesia, Jakarta;
[3]Faculty of Medicine, Universitas Muhammadiyah Palembang, Palembang;
[4]Faculty of Medicine, Universitas Padjadjaran/Hasan Sadikin General Hospital, Bandung;
[5]Department of Otorhinolaryngology, Head and Neck Surgery, Harapan Kita National Women and Children Health Center, Jakarta;
[6]Department of Anatomical Pathology, Faculty of Medicine, Hasanuddin University, Makassar;
[7]Department of Anatomical Pathology, Faculty of Medicine, Andalas University, Padang;
[8]Department of Anatomical Pathology, Faculty of Medicine, Universitas Brawijaya/RSUD dr. Saiful Anwar, Malang;
[9]Department of Anatomical Pathology, Faculty of Medicine and Health Science, Universitas Jambi, Jambi;
[10]Department of Anatomical Pathology, Faculty of Medicine, Universitas Padjadjaran/Hasan Sadikin General Hospital, Bandung;
[11]Department of Anatomical Pathology, Faculty of Medicine, Universitas Diponegoro, Semarang;
[12]Department of Anatomical Pathology, Faculty of Medicine, University of Lampung, Lampung;
[13]Department of Anatomical Pathology, Faculty of Medicine, University of Sriwijaya, Palembang;
[14]Department of Anatomical Pathology, Labuha Hospital, South Halmahera;
[15]Department of Anatomical Pathology, Faculty of Medicine, Universitas Udayana, Prof. Dr. I.G.N.G. Ngoerah Hospital, Denpasar;
[16]Department of Anatomical Pathology, Faculty of Medicine, Universitas Airlangga/Dr Soetomo Academic Hospital, Surabaya;
[17]Department of Anatomical Pathology, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada/UGM Academic Hospital, Yogyakarta;
[18]Department of Anatomical Pathology, Dr. Ben Mboi Hospital, Kupang;
[19]Kanujoso Djatiwibowo Hospital, Balikpapan, Indonesia;
[20]Department of Pathology and Thyroid Disease Center, Izumi City General Hospital, Izumi, Japan

Background: The diagnosis of thyroid neoplasms necessitates the identification of distinct histological features. Various education/hospital centers located in cities across Indonesia likely result in discordances among pathologists when diagnosing thyroid neoplasms. **Methods:** This study examined the concordance among Indonesian pathologists in assessing nuclear features and capsular and vascular invasion of thyroid tumors. Fifteen pathologists from different centers independently assessed the same 14 digital slides of thyroid tumor specimens. All the specimens were thyroid neoplasms with known *BRAF*V600E and *RAS* mutational status, from a single center. We evaluated the pre- and post-training agreement using the Fleiss kappa. The significance of the training was evaluated using a paired T-test. **Results:** Baseline agreement on nuclear features was slight to fair based on a 3-point scoring system (k=0.14 to 0.28) and poor to fair based on an eight-point system (k=–0.02 to 0.24). Agreements on vascular (κ=0.35) and capsular invasion (κ=0.27) were fair, whereas the estimated molecular type showed substantial agreement (κ=0.74). Following the training, agreement using the eight-point system significantly improved (p=0.001). **Conclusions:** The level of concordance among Indonesian pathologists in diagnosing thyroid neoplasm was relatively poor. Consensus in pathology assessment requires ongoing collaboration and education to refine diagnostic criteria.

Key Words: Thyroid neoplasms; Observer variation; Papillary thyroid cancer

Received: June 23, 2024 **Revised:** July 18, 2024 **Accepted:** July 24, 2024
Corresponding Author: Agnes Stephanie Harahap, Department of Anatomical Pathology, Faculty of Medicine, Universitas Indonesia/Dr. Cipto Mangunkusumo Hospital, Jl. Salemba Raya No. 6, Jakarta, 14320, Indonesia
Tel: +62-8-18765563, Fax: +62-21-3912477, E-mail: agnes.stephanie01@ui.ac.id

pISSN 2383-7837
eISSN 2383-7845

In 2022, thyroid neoplasm ranked as the seventh most prevalent cancer worldwide, with approximately 80%–90% of cases being papillary thyroid carcinoma (PTC) [1,2]. Histologically, PTC is a well-differentiated thyroid malignancy originating from follicular cells of the thyroid gland. The incidence of PTC has increased over recent decades [3,4]; however, the reasons for this increase have been the subject of considerable debate among global experts [5]. The major risk factors for PTC include ionizing radiation and obesity, the incidences of which have also continued to increase within the last decade [1]. However, the increasing attention to thyroid examination may also lead to greater detection of subclinical papillary carcinoma, as seen in developed countries such as South Korea and the United States [1]. Despite the increased incidence, mortality rates have remained unchanged, suggesting a favorable prognosis with a 10-year cause-specific survival rate in 73% of patients [3,4].

Establishing the diagnosis of PTC involves assessment of the gross pathology and histology of the thyroid gland. The gross appearance of the tumor may manifest as either solid nodules or cystic structures within the thyroid gland. Definitive diagnosis of PTC requires microscopic identification of papillary, follicular, and/or solid architectural structures, specific nuclear features, and the determination of the infiltrative pattern of the tumor [2]. Under the current World Health Organization (WHO) classification, various tumor diagnoses are encompassed under 'encapsulated follicular-patterned thyroid lesions' depending on the nuclear features and capsular and vascular invasion [2]. Reproducibility, particularly concerning the nuclear features of PTC and the presence of capsular and vascular invasion, was a significant concern among experts. Previous studies have shown considerable variation among pathologists when diagnosing the nuclear features of PTC, especially in tumors showing encapsulated follicular patterns [6-10]. Other studies have also highlighted challenges in the evaluation of thyroid tumor capsular invasion and minimal extrathyroidal invasion [11,12]. This variability encompasses both interobserver variability, which refers to differences between experts, and intra-observer variability, which pertains to variations within the same expert. The absence of well-defined evidence-based diagnostic criteria is the primary reason for this variability.

Various nuclear features have been utilized to diagnose PTC, including identification of tumor cells with high nuclear-to-cytoplasm ratio; nuclear enlargement; nuclear overlap, crowding, and elongation; irregular nuclear contours; intranuclear cytoplasmic inclusions; chromatin clearing; and multiple nucleoli located near the cell membrane [7]. To reduce variation in histological assessments, the WHO has adopted the three-point scoring system originally proposed by Nikiforov et al. [13] to evaluate nuclear features of PTC with low subjectivity. The scoring system consists of three measures: cell nucleus size and shape (enlarged, elongated, and crowded), nuclear membrane features (irregular nuclear contour, nuclear grooves, and nuclear pseudoinclusion), and chromatin characteristics (clear chromatin, chromatin margination to the cell membrane, and glassy nuclei). Each criterion contributes a score of 1, resulting in a total nuclear score range of 0–3. A nuclear score of 0–1 does not meet the criteria for typical PTC, whereas a score of 2–3 is considered typical. In 2018, the Asian Thyroid Working Group proposed a more detailed eight-point scoring system to evaluate the nuclear features of PTC [6]. In a study by Jung et al. [14], the eight-point scoring system was able to distinguish *BRAF*-like tumors from *RAS*-like tumors based on nuclear pseudo-inclusion and a high nuclear score.

Indonesia comprises more than 17,000 widely dispersed islands, which are organized into 38 provinces. Presently, Indonesia is home to approximately 860 pathologists who exhibit a wide array of educational backgrounds and professional experience. The presence of various medical education centers in Indonesia may also contribute to a higher likelihood of interobserver disagreement in the diagnosis of certain diseases.

To date, there have been no interobserver studies aimed at evaluating the extent of concordance among Indonesian pathologists in the assessment of the nuclear features and capsular and vascular invasion of PTC. The primary objective of this study was to assess the level of agreement among pathologists from various centers in Indonesia regarding the diagnosis of nuclear features as well as capsular/vascular invasion and the presumed molecular type of PTC cases.

MATERIALS AND METHODS

Study design

We conducted an interobserver agreement study in which 15 pathologists from multiple educational/hospital centers in Indonesia were recruited to perform an independent histology assessment of thyroid neoplasm cases. As depicted in Fig. 1, the participating pathologists were from Padang, Palembang, Bandar Lampung, Jambi, DKI Jakarta, Bandung, Yogyakarta, Surabaya, Semarang, Malang, Bali, Balikpapan, Makassar, Kupang, and South Halmahera.

This study consists of two rounds of histological assessment with two online training sessions in between. For histological

 https://doi.org/10.4132/jptm.2024.07.25

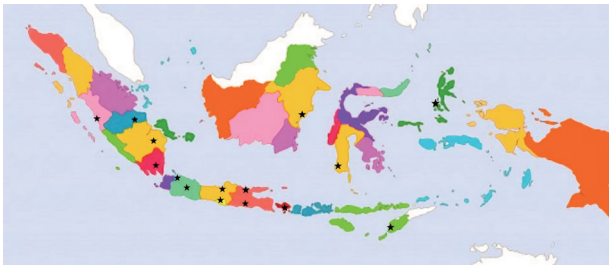

Fig. 1. The geographical distribution of participating pathologists. Each main island of Indonesia had at least one participating pathologist, with the largest number coming from Java and Sumatra.

assessment, the pathologists received digital slides of thyroid tumor specimens along with a questionnaire in which to record their histology conclusions. Basic demographic data about the pathologist's age; sex; numbers of years in pathology practice, in a teaching hospital, and/or in thyroid consultancy; number of thyroid cases diagnosed per year; history of study abroad; the pathologist's experience in diagnosing non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) in pathology practice; the routine acquisition of clinical and radiology information; and the pathologist's prior experience with digital pathology slides were included in the initial questionnaire during the first round. In addition, pathology data including the method used for tumor capsular sectioning and the number of blocks submitted for a 4-cm thyroid nodule were also analyzed.

After the first round, all pathologists received a learning module (Supplementary Materials) and attended two online training sessions directed by the authors. To evaluate the impact of the training, a second-round assessment was performed two months after the first.

Digital slides of tumor specimens

The tumor specimens used in this study were thyroid neoplasms collected from Dr. Cipto Mangunkusumo Hospital, Jakarta, Indonesia. The histologic slides were stained with hematoxylin and eosin and were scanned using Aperio software (Leica Biosystems, Buffalo Grove, IL, USA) and converted into digital slides accessible for online viewing. Both rounds of assessment consisted of 14 cases of thyroid neoplasms. The second round comprised seven cases previously examined in the first round and seven new cases. We randomly mixed the sample cases in the second round to prevent memory bias. We determined the mutational status of *BRAF*V600E and *RAS* for all specimens following a previous procedure [15]. Seven of each *BRAF*V600E and *RAS* mutations were identified in each round.

This study evaluated the pathologists' diagnoses of thyroid neoplasm, particularly their evaluation of the nuclear features of PTC. We asked each pathologist to assess the nuclear features based on the two nuclear scoring systems. As noted above, the three-point scoring system developed by Nikiforov et al. [13] encompasses three key parameters of nuclear size and shape, membrane abnormalities, and chromatin characteristics (Table 1). The evaluation of nuclear features was enhanced by a comprehensive eight-point scoring system modified from the schemes of Liu et al. [6] and Jung et al. [14]. The eight-point scoring system used in the current study encompassed the assessment of nuclear enlargement, nuclear crowding, nuclear elongation, irregular nuclear contour, nuclear grooves, nuclear pseudo-inclusion, chromatin clearing, and PTC-nuclear feature distribution. Specific scores were assigned to each of these features, as outlined in Table 2. The nuclear examination was conducted using adjacent benign follicular thyroid tissue as a comparative reference.

Table 1. Three-point nuclear score system

Nuclear feature	Score
Nuclear size and shape	
Nuclear enlargement	0: Absent or only slightly expressed
Nuclear crowding/overlapping	1: Present or well developed
Nuclear elongation	
Membrane irregularities	
Irregular membrane contour	0: Absent or only slightly expressed
Nuclear grooves	1: Present or well developed
Nuclear pseudoinclusions	
Chromatin characteristics	
Chromatin clearing	0: Absent or only slightly expressed
Margination of chromatin to membrane	1: Present or well developed
Glassy nuclei	
Total score	0–1: Not diagnostic for PTC nuclei
	2–3: Diagnostic for PTC nuclei

PTC, papillary thyroid carcinoma.

In addition to the evaluation of nuclear features, assessment of capsular and vascular invasion (categorized as invasive, non-invasive, or equivocal), molecular status (classified as BRAF-like or RAS-like), and whether the case was straightforward or doubtful was also completed.

Statistical analyses

The pathologists' demographic information was presented in the form of categorical data, displayed as frequencies and percentages. The levels of interobserver agreement in the first and second rounds were analyzed using Fleiss kappa statistical analysis. The interpretation of kappa's statistical result was as follows: 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–0.99, almost perfect agreement. A kappa value

Table 2. Eight-point nuclear scoring system

Nuclear feature	Score
Nuclear enlargement	0: Absent 1: Present in < 10% of tumor cells 2: Present in ≥ 10% of tumor cells
Nuclear crowding/ overlapping	0: Absent 1: Present in < 10% of tumor cells 2: Present in ≥ 10% of tumor cells
Nuclear elongation	0: Absent 1: Present in < 10% of tumor cells 2: Present in ≥ 10% of tumor cells
Irregular membrane contour	0: Absent 1: Present in < 10% of tumor cells 2: Present in ≥ 10% of tumor cells
Nuclear grooves	0: Absent 1: Present in < 10% of tumor cells 2: Present in ≥ 10% of tumor cells
Nuclear pseudoinclusions	0: Absent 1: Present
Chromatin clearing	0: Absent 1: Present in < 10% of tumor cells 2: Present in ≥ 10% of tumor cells
PTC-nuclear features distribution	0: Focal 1: Diffuse
Total score	0–14

PTC, papillary thyroid carcinoma.

of 1 indicated perfect agreement, while values < 0 indicated pure chance alone. A stratified analysis based on endocrine expertise was also performed. All statistical analyses were performed using SPSS software ver. 20 (IBM Corp., Armonk, NY, USA).

RESULTS

The average age of pathologists in this study was 46 years, and the majority was female. Most participating pathologists were based in teaching hospitals, with an average of 10 years of practical experience. Nine respondents were pathologists who had passed national board examinations with a subspecialty in endocrinology. The number of thyroid cases diagnosed by the respondents ranged from 21 to 1,200 per year, with a median of 110 cases and interquartile range of 50–217 cases (Table 3). The tumor sampling techniques varied between the pathologists; some pathologists sectioned the entire capsule of follicular-patterned tumors, while all other pathologists sampled < 8 blocks. Most of the pathologists were familiar with digital pathology.

The sample cases exhibited varying degrees of ambiguity (Fig. 2), which consequently influenced the interobserver agreement. The greatest level of agreement among pathologists was reported for a case with a positive *RAS* mutation, for which most pathologists provided a total nuclear score of 1. The assessment for other histological parameters such as capsular invasion

Table 3. Demographics of participating pathologists

Characteristic	No. (%)
Age (yr)	
Mean ± SD	45.7 ± 5.1
31–40	1 (6.7)
41–50	12 (80.0)
51–60	2 (13.3)
Sex	
Male	3 (20.0)
Female	12 (80.0)
Years of practice	
Mean ± SD	10.3 ± 3.9
0–5	2 (13.3)
6–10	7 (46.7)
11–20	6 (40.0)
Teaching hospital	
Yes	12 (80.0)
No	3 (20.0)
Pathology consultant	
Yes	9 (60.0)
No	6 (40.0)
Thyroid case/yr	
Median (min–max)	110 (21–1,200)
IQR (25–75)	50–217
Study pathology abroad	
Yes	5 (33.3)
No	10 (66.7)
Length of study (mo)	
Mean ± SD	8.2 ± 9.9
1–6	3 (60.0)
7–12	1 (20.0)
13–18	0
19–24	1 (20.0)
Topic/field of the study	
Bone marrow pathology	1 (20.0)
Lymphoma	1 (20.0)
Molecular pathology	2 (40.0)
Vascular	1 (20.0)
Section entire tumor capsule	
Yes	7 (46.7)
No	8 (53.3)
Blocks sample in a 4 cm thyroid nodule	
< 5 blocks	4 (50.0)
5–8 blocks	4 (50.0)
> 8 blocks	0
Diagnose NIFTP in pathology practice	
Yes	15 (100)
No	0
Receive clinical and radiological information in routine practice	
Yes	6 (40.0)
Sometimes	8 (53.3)
No	1 (6.7)
First time using digital pathology	
Yes	4 (26.7)
No	11 (73.3)

SD, standard deviation; IQR, interquartile range; NIFTP, non-invasive follicular thyroid neoplasm with papillary-like nuclear features.

https://doi.org/10.4132/jptm.2024.07.25

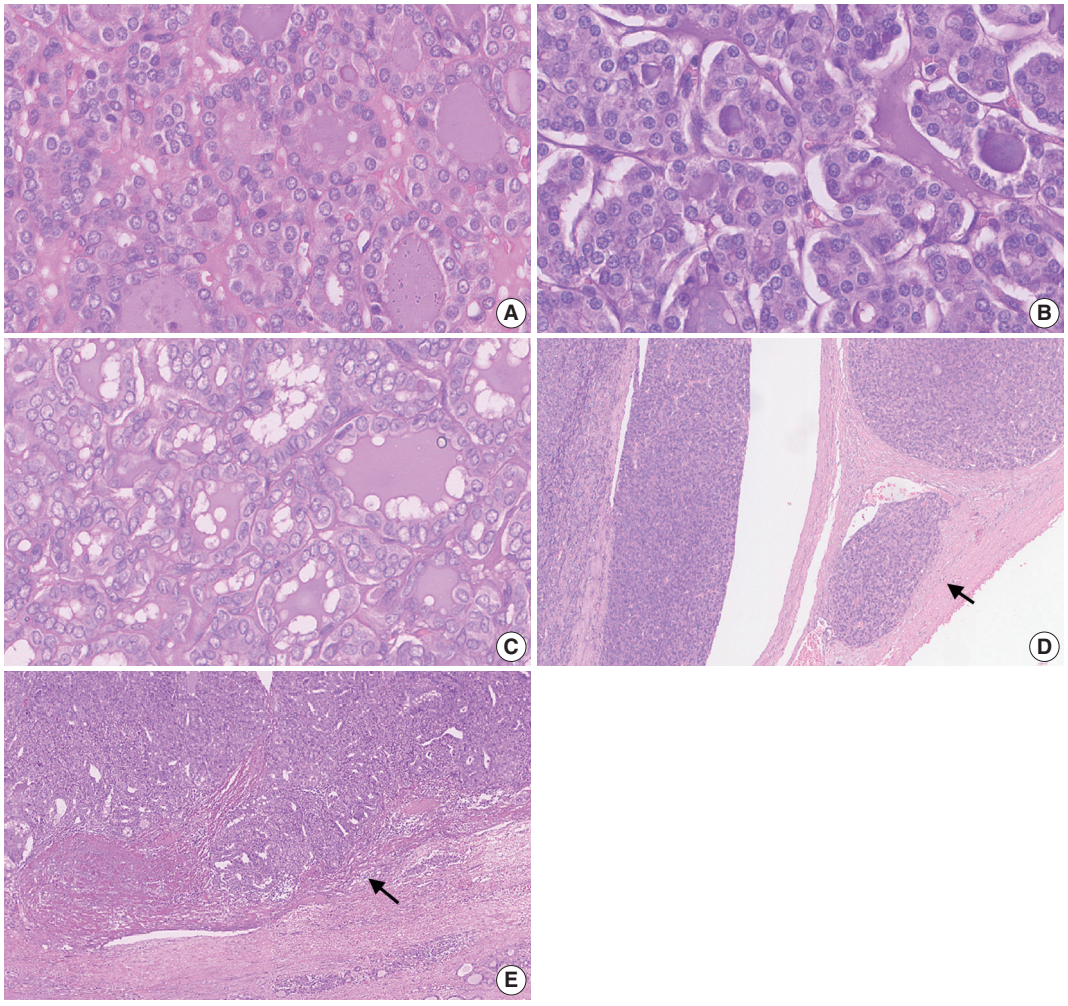

Fig. 2. (A) The assessment of the microscopic appearance of this *NRAS*Q61R-mutated case led to the most significant interobserver disagreement when using a three-point scoring system, primarily due to the ambiguous nature of the nuclear features. (B) The assessment of this *HRAS*Q61R mutated-case led to the strongest interobserver agreement when assessed using a three-point system. Most pathologists assigned a total nuclear score of 1 for this case. (C) A *BRAF*V600E-mutated papillary thyroid carcinoma case showing marked nuclear alterations such as nuclear crowding, an irregular nuclear membrane, chromatin clearing, and nuclear grooves, with most pathologists agreeing on a total score of 3. (D) All pathologists noted vascular invasion (arrow). (E) This case demonstrated the greatest level of interobserver disagreement regarding capsular invasion, with some pathologists indicating "no" and others indicating "equivocal" (arrow).

also varied between the pathologists.

Pathologist agreement regarding the assessment of nuclear features

In the first round of assessments using the three-point system, the Fleiss kappa values were as follows: nuclear size and shape (κ = 0.14), chromatin features (κ = 0.21), and membrane irregularities (κ = 0.28). These findings indicated slight to fair agreement among pathologists when evaluating each nuclear feature. When evaluating the total score for nuclear features using the three-point system, only slight agreement between the pathologists assessments was evident (κ = 0.17). Using the eight-point nuclear scoring system, the Fleiss kappa values represented poor to fair agreement (Table 4), with the highest kappa values obtained for the assessment of nuclear pseudo-inclusion (κ = 0.24) and nuclear elongation (0.20). The lowest kappa value was obtained for the assessment of PTC nuclear features distribution (κ = –0.02) and nuclear crowding/overlapping (κ = 0.01). The concordance for the total score of nuclear features based on the eight-point system was poor (κ = –0.01).

Although not statistically significant (p = .052), there was an improvement in the average kappa agreement of the three-point nuclear score system following the training (Table 5). In the second (post-training) round, the pathologists displayed fair

Table 4. Agreement regarding nuclear features assessment using the three-point and eight-point scoring systems

Nuclear feature	Round one				Round two			
	Kappa	p-value	95% CI	Strength	Kappa	p-value	95% CI	Strength
Three-point scoring system								
Nuclear size and shape	0.14	<.001	0.14 to 0.14	Slight	0.30	<.001	0.30 to 0.30	Fair
Membrane irregularities	0.28	<.001	0.28 to 0.28	Fair	0.57	<.001	0.57 to 0.58	Moderate
Chromatin features	0.21	<.001	0.20 to 0.21	Fair	0.59	<.001	0.59 to 0.59	Moderate
Total score	0.17	<.001	0.17 to 0.17	Slight	0.39	<.001	0.39 to 0.39	Fair
Eight-point scoring system								
Nuclear enlargement	0.09	.001	0.08 to 0.09	Slight	0.36	<.001	0.36 to 0.36	Fair
Nuclear crowding/overlapping	0.01	.987	–0.01 to 0.01	Slight	0.38	<.001	0.38 to 0.38	Fair
Nuclear elongation	0.20	<.001	0.20 to 0.20	Slight	0.37	<.001	0.37 to 0.38	Fair
Irregular membrane contour	0.13	<.001	0.13 to 0.13	Slight	0.48	<.001	0.48 to 0.48	Moderate
Nuclear grooves	0.07	.005	0.07 to 0.08	Slight	0.19	<.001	0.19 to 0.20	Slight
Nuclear pseudo-inclusion	0.24	<.001	0.24 to 0.24	Fair	0.35	<.001	0.35 to 0.36	Fair
PTC-nuclear features distribution	–0.02	.481	–0.02 to –0.02	Poor	0.23	<.001	0.23 to 0.23	Fair
Chromatin clearing	0.12	<.001	0.12 to 0.12	Slight	0.38	<.001	0.38 to 0.39	Fair
Total score	–0.01	.611	–0.01 to –0.01	Poor	0.05	<.001	0.04 to 0.05	Slight

CI, confidence interval; PTC, papillary thyroid carcinoma.

Table 5. Paired T-test analysis on the average kappa value before and after training

Variable	Round	Mean kappa	95% CI	p-value
Three-point score system	First	0.21	–0.57 to 0.01	.052
	Second	0.49		
Eight-point score system	First	0.10	–0.32 to –0.16	.001
	Second	0.34		

CI, confidence interval.

to moderate agreement when evaluating the nuclear size and shape (κ = 0.30), membrane irregularities (κ = 0.57), and chromatin features (κ = 0.59). When evaluating the total score of nuclear features based on the three-point system, the pathologists displayed a fair level of agreement (κ = 0.39). Furthermore, there was a significant improvement in the agreement when using the eight-point nuclear scoring system (p = .001). The pathologists displayed fair agreement when assessing most nuclear features and moderate agreement when evaluating irregularities of the membrane contour (κ = 0.48). The only feature with slight agreement was nuclear grooves (κ = 0.19). The agreement for the total nuclear score based on the eight-point system was slight (κ = 0.05). As displayed in Fig. 3, total nuclear scores for both the three- and eight-point systems were lower in the second round than in the first round.

The study also compared the assessments of the 15 pathologists in the two rounds of slide assessments (Fig. 4). Four pathologists (Nos. 4, 7, 9, and 15) demonstrated uniformity in their average nuclear scores with the three-point assessment between the two rounds. Upon analysis of the scores obtained using the eight-point scoring system, two pathologists (Nos. 2 and 7) demonstrated marked differences in nuclear scoring assessment compared to the initial round, while only three pathologists (Nos. 4, 9, and 13) showed minimal alterations in the evaluative outcomes between the two rounds.

Stratified analysis based on the expertise of pathologists

Using the three-point system, the interobserver agreement for nuclear features was higher among endocrine experts compared to non-endocrine experts (Fig. 5). The greatest agreement among the endocrine experts was obtained when assessing chromatin features (κ = 0.52), followed by membrane irregularities (κ = 0.27) and nuclear size and shape (κ = 0.22). In comparison, non-endocrine experts displayed fair agreement for membrane irregularities (κ = 0.21) and slight agreement for both chromatin features (κ = 0.10) and nuclear size and shape (κ = 0.08). The endocrine experts displayed fair agreement for the total nuclear score of the three-point system (κ = 0.26) relative to the slight agreement seen among the non-experts (κ = 0.11).

Using the eight-point nuclear score system, both the endocrine and non-endocrine experts showed variation in their level of concordance (Fig. 5). The endocrine experts displayed better agreement regarding nuclear pseudo-inclusion (κ = 0.31) relative to the non-endocrine experts (κ = 0.18). However, the non-endocrine experts displayed more concordance when assessing other nuclear features, such as nuclear elongation, nuclear enlargement, irregular membrane contour, and chromatin clearing.

After the training, there was significant improvement in the agreement between the endocrine and non-endocrine experts when assessing nuclear features (Fig. 6). In the second (post-

https://doi.org/10.4132/jptm.2024.07.25

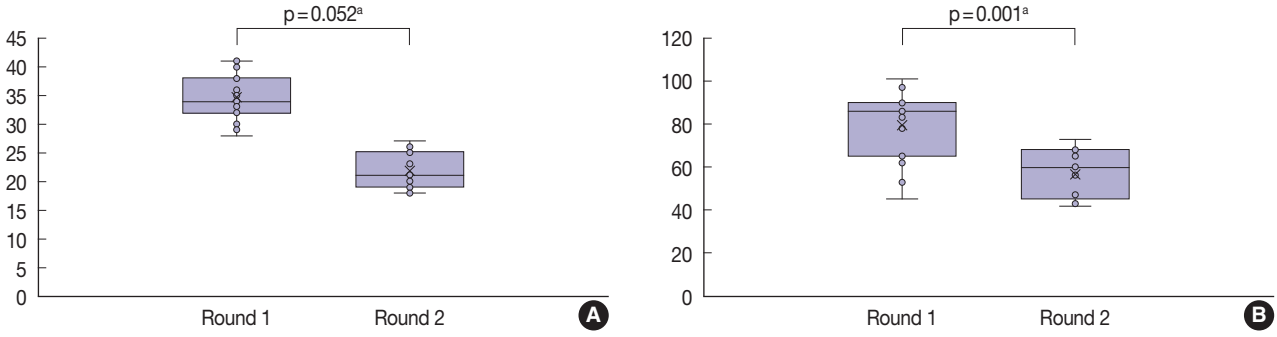

Fig. 3. Box plots of total nuclear scores using the three-point scoring system (A) and the eight-point scoring system (B). As shown by the vertical axis, the total nuclear score after training was lower and showed a narrower range compared to the first round. Using the eight-point score system, there was a significant difference between the level of agreement in the first and second rounds (p=.001). [a]The p-value is obtained from the mean calculation of Kappa value using a paired T-test.

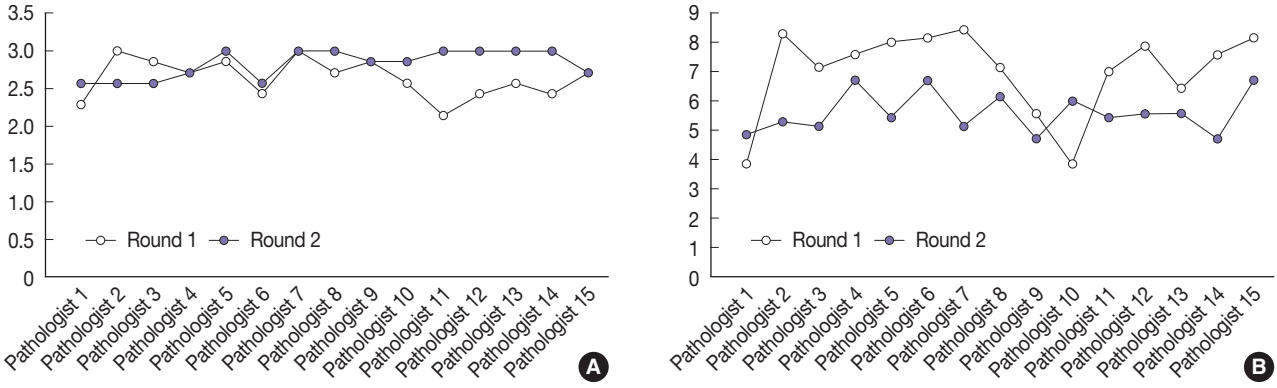

Fig. 4. Comparative analysis of the average sum of nuclear scores provided by 15 pathologists across two rounds when using the three-point scoring system (A) and eight-point scoring system (B).

training) round, the experts displayed moderate agreement when assessing each nuclear feature using the three-point system: κ = 0.52 for membrane irregularities; κ = 0.44 for nuclear size and shape; and κ = 0.44 for chromatin features. Furthermore, the non-endocrine experts displayed substantial agreement when assessing chromatin features (κ = 0.69), moderate agreement for membrane irregularities (κ = 0.57), and fair agreement for nuclear size and shape (κ = 0.20). Both the endocrine experts (κ = 0.34) and non-experts (κ = 0.39) reached fair agreement when assessing the total nuclear score using the three-point system. Using the eight-point system, the agreement level showed considerable variation, as displayed in Fig. 6. The endocrine experts showed better agreement when assessing nuclear elongation, nuclear crowding/overlapping, nuclear pseudo-inclusion, and chromatin clearing compared to the non-experts.

Pathologists' agreement for other histological parameters

This study also analyzed interobserver agreement for the assessment of additional histological parameters, including the presumed molecular type and the presence of vascular and/or capsular invasion (Table 6). The kappa statistics indicated substantial agreement for all pathologists (κ = 0.74) regarding the presumed genetic type, with an almost perfect agreement level among endocrine experts. The accuracy of the molecular type identification ranged from 67%–100%, with 10 pathologists assigning 100% correct answers. The assessment of vascular invasion reached fair agreement among all pathologists (κ = 0.35). A stratified analysis displayed better agreement (k = 0.53) for the assessment of vascular invasion among the non-endocrine experts compared to the experts. The capsular invasion agreement was fair among all groups.

DISCUSSION

Obtaining consistency in thyroid neoplasm diagnosis presents a challenge to pathologists across the globe. Variation between assessments performed by individual pathologists may arise due to the application of different standard criteria in different geo-

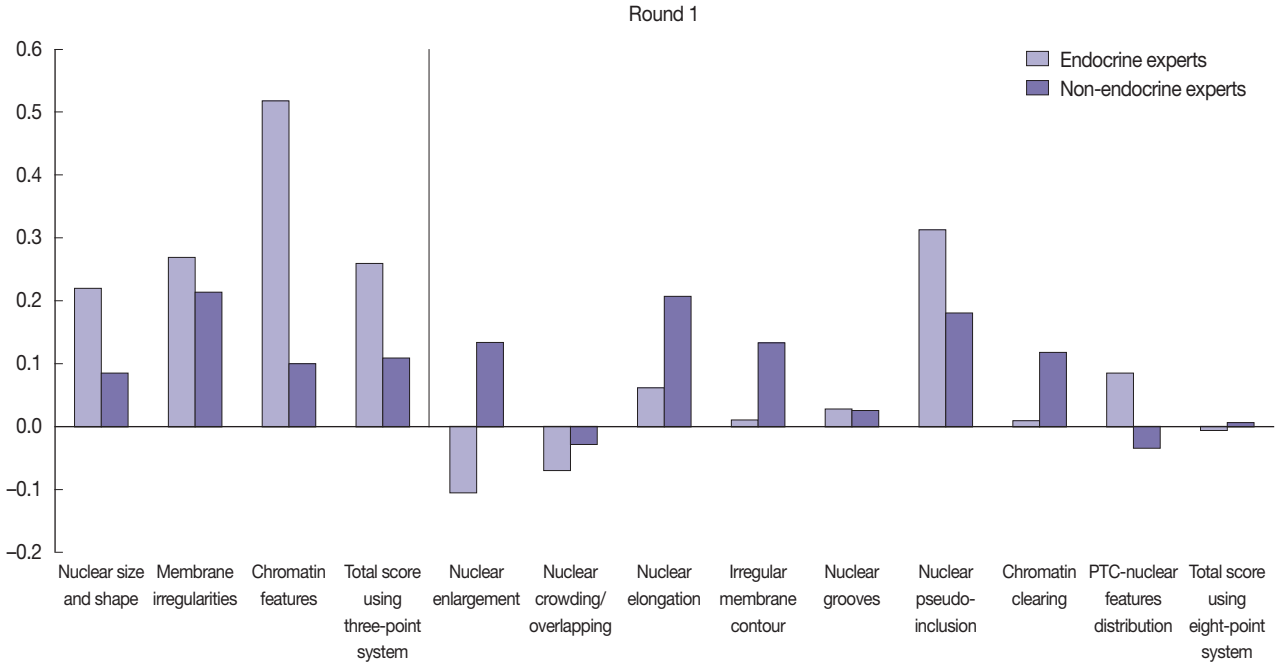

Fig. 5. Comparison of kappa values for nuclear features between endocrine and non-endocrine experts in the first round. PTC, papillary thyroid carcinoma.

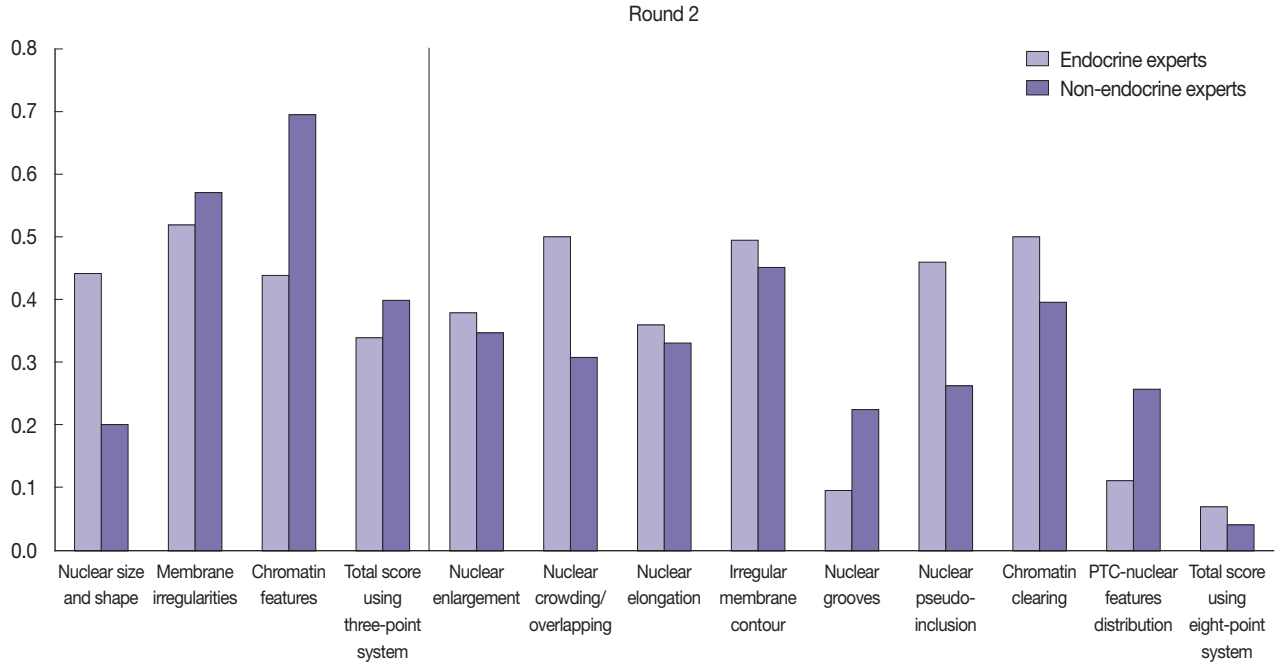

Fig. 6. Comparison of kappa values for nuclear features between endocrine and non-endocrine experts in the second round. PTC, papillary thyroid carcinoma.

graphic regions [9,10]. Pathologists in developed countries often employ ancillary tools such as immunohistochemistry and molecular examinations to assist in the establishment of a diagnosis. Although such methods are widely accepted, it is crucial to understand that the final diagnosis of thyroid neoplasm relies on the assessment of morphological features [16]. Our study confirmed that pathologists in Indonesia, influenced by their educational background, experiences, and viewpoints, demonstrated only moderate agreement when applying the standardized diagnostic criteria. We assessed concordance among pathologists

 https://doi.org/10.4132/jptm.2024.07.25

Table 6. Agreement regarding the assessment of other histological parameters

	Molecular type				Vascular invasion				Capsular invasion			
	Kappa	p-value	95% CI	Strength	Kappa	p-value	95% CI	Strength	Kappa	p-value	95% CI	Strength
All pathologists	0.74	<.001	0.74–0.74	Substantial	0.35	<.001	0.35–0.36	Fair	0.27	<.001	0.29–0.27	Fair
Endocrine experts	0.84	<.001	0.84–0.85	Almost perfect	0.21	<.001	0.21–0.22	Fair	0.26	<.001	0.26–0.27	Fair
Non-endocrine experts	0.69	<.001	0.68–0.69	Substantial	0.53	<.001	0.52–0.52	Moderate	0.25	<.001	0.25–0.26	Fair

CI, confidence interval.

from several cities in Indonesia with varied educational backgrounds and working environments when assessing the histopathological features of thyroid neoplasms. The discordances highlight the need for clear and uniform standards in pathological evaluation to reduce variability in interpretations.

The nuclear features of thyroid neoplasms range widely from nuclei with minimal atypia to prominent nuclear changes. The heterogeneity of the nuclear morphology is due in part to the presence of molecular alterations such as *BRAF*V600E and *RAS* mutations. Tumors with *BRAF*V600E mutations have more prominent PTC nuclear features, while those with *RAS* mutation show less noticeable nuclear features [14]. Additional genetic alterations may also contribute to the typical morphological features of PTC, including translocation of the *NTRK* gene, *THADA-IGF2BP3* gene fusion, and *DICER1* mutation [17-19]. The three-point nuclear scoring system is currently used by experts to diagnose NIFTP and papillary carcinoma. A total nuclear score of 0–1 favors a diagnosis of benign tumors while a total nuclear score of 2–3 favors NIFTP or carcinoma. This study found only slight to fair baseline concordance among Indonesian pathologists when assessing nuclear features using the three-point system. This result indicated lower concordance than was reported in an earlier validation study of the three-point nuclear system that showed good to substantial concordance among pathologists across California, the UK, and Japan [7]. Nonetheless, consistent with this earlier study, nuclear size and shape were the criteria with the lowest concordance, whereas membrane irregularity was the criterion with the highest agreement level [7].

Although the three-point system acts to standardize the diagnosis of nuclear features in thyroid tumors, continuous refinement is needed to accommodate the complexity of pathological features while ensuring an accurate and consistent diagnosis [7]. We also evaluated concordance among pathologists when evaluating nuclear features using the eight-point system. The baseline concordance among pathologists ranged widely from poor to fair. The pathologists demonstrated the greatest agreement for nuclear pseudo-inclusion. The three-point scoring system demonstrated better agreement than the eight-point scoring system, suggesting that its simplicity may facilitate more consistent assessments.

The training was significantly effective in enhancing the level of agreement among participating pathologists, particularly when using the eight-point nuclear scoring system. Following the training, moderate agreement was obtained for assessment of membrane irregularities, chromatin features, and irregular membrane contours. Interestingly, the feature of nuclear grooves yielded the greatest discordance, even after training. This finding contrasts with the results from a study by Elsheikh et al. [9], which identified nuclear grooves as having the second-highest level of agreement after nuclear clearing. Despite recent enhancements, the concordance achieved with the scoring system remains suboptimal, potentially attributable to variability in the experience of and thyroid case volumes among pathologists. The study conducted by Thompson et al. [7] demonstrated a minor trend in which pathologists practicing in the UK exhibited slightly lower accuracy than those in California and Japan.

The comparative analysis of performance scores for the 15 pathologists across two evaluation rounds provided insights into the dynamic nature of diagnostic assessments. This analysis of this line chart highlighted significant fluctuations in individual pathologist scores within each round, suggesting variability in the application of the scoring systems. These observations underscore the complexities of achieving uniformity in pathological assessment and highlight the need for ongoing training and calibration among pathologists to enhance the reliability of the processes used to diagnose thyroid neoplasms.

Another important finding was the correlation between the clinician's area of expertise and the level of diagnostic agreement. Our findings suggest that endocrine experts often reach consensus on cellular characteristics, while non-endocrine experts demonstrate consistency in other aspects. According to Farmer et al. [20], expert input often significantly alters the initial diagnosis, with a very low kappa coefficient indicating substantial differences between assessments by expert and non-expert pathologists. Kerkhof et al. [21] reported moderate agreement among pathologists with varying levels of expertise when evaluating

dysplasia, with greater concordance achieved when assessing more advanced cases. Notably, Thompson et al. [7] found that a uniform scoring system fosters agreement among pathologists in different subspecialties when diagnosing papillary carcinoma. Together, these findings suggest that, while expertise is irreplaceable, particularly in complex cases, the definition and adoption of clear, structured diagnostic criteria may provide common ground upon which to establish diagnostic objectivity and clarity, and interobserver agreement should improve.

Additional histological features contributing to the diagnostic challenges presented by PTC include capsular and/or vascular invasion. The categories of capsular invasion span non-invasive, questionable invasion, and unequivocally invasive forms. Zhu et al. [11] reported fair agreement among thyroid pathologists for non-invasive and invasive tumors, whereas questionable invasion elicited poor agreement, indicating that borderline tumors pose a significant challenge in the practice of thyroid pathology. The present study, which involved the evaluation of a mixture of encapsulated and invasive cases, resulted in fair agreement overall in determining capsular invasion. The identification of vascular invasion holds critical importance not only in establishing a definitive diagnosis, but also in prognostication and assessing the aggressiveness of the disease [22]. This study demonstrated fair to moderate agreement among pathologists when evaluating vascular invasion. The assessment of vascular invasion is complicated by several factors, including the presence of clefts that mimic vessels and artifacts related to floating tumor cells. Some immunohistochemistry markers (such as CD 31 or ERG) might help to evaluate this feature more robustly but were not used in this study.

The potential to utilize histological assessment to predict genetic alteration of thyroid tumors has been emphasized by several experts [15,19]. In the present study, the agreement among pathologists regarding the identification of tumors with *BRAF*V600E or *RAS* mutations ranged from substantial to nearly perfect. Ten pathologists involved in this study demonstrated 100% accuracy. The three-point scoring system exhibited 85%–94% accuracy for prediction of molecular alterations in NIFTP [19]. Notably, nuclear pseudo-inclusion is believed to be a distinctive feature of *BRAF*V600E-mutated PTC.

The major limitation of this study was the heterogeneous background and broad experience among pathologists. The small number of pathologists limited the generalizability of the findings to the large population of Indonesian pathologists.

Following the training, the overall agreement among pathologists when assessing nuclear features using the three- and eight-point scoring systems was slight to moderate. There was fair agreement regarding the identification of capsular invasion, whereas fair to moderate agreement was achieved for vascular invasion. The level of agreement was greatest for presumed genetic type. Despite enhancements to the scoring systems, the level of agreement attained in this study remained below optimal, possibly due to the diverse backgrounds of the individual pathologists. The pursuit of national standardization in pathological assessment, and especially with respect to the evaluation of nuclear features, is an ongoing effort that requires collaboration among pathologists to refine the applicable criteria, enhance training, and ultimately improve the reliability and validity of diagnostic processes in the field of thyroid pathology.

Supplementary Information

The Data Supplement is available with this article at https://doi.org/10.4132/jptm.2024.07.25.

Ethics Statement

This study was approved by the Ethics Committee of the Faculty of Medicine Universitas Indonesia – Dr. Cipto Mangunkusumo Hospital, under protocol number KET-620/UN2.F1/ETIK/PPM.00.02.2023. The requirement for informed consent was waived by the committee under protocol number ND-0313/UN2.F1.DEPT.27/PPM.00.02/2023.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Code Availability

Not applicable.

ORCID

Agnes Stephanie Harahap	https://orcid.org/0000-0001-8920-7873
Mutiah Mutmainnah	https://orcid.org/0009-0000-6036-0647
Maria Francisca Ham	https://orcid.org/0000-0002-7915-5536
Dina Khoirunnisa	https://orcid.org/0009-0006-1801-3223
Abdillah Hasbi Assadyk	https://orcid.org/0009-0000-8944-4537
Husni Cangara	https://orcid.org/0000-0002-5160-8265
Aswiyanti Asri	https://orcid.org/0000-0001-7073-3240
Diah Prabawati Retnani	https://orcid.org/0000-0003-4300-0855
Fairuz Quzwain	https://orcid.org/0000-0002-6871-1189
Hasrayati Agustina	https://orcid.org/0000-0001-8817-6753
Hermawan Istiadi	https://orcid.org/0009-0001-2367-9218
Indri Windarti	https://orcid.org/0000-0001-8594-7137
Krisna Murti	https://orcid.org/0000-0001-6733-2323
Muhammad Takbir	https://orcid.org/0009-0005-8390-974X
Ni Made Mahastuti	https://orcid.org/0009-0009-9029-5423
Nila Kurniasari	https://orcid.org/0000-0001-5584-2705
Nungki Anggorowati	https://orcid.org/0000-0002-3268-5492
Pamela Abineno	https://orcid.org/0009-0003-2400-1745
Yulita Pundewi Setyorini	https://orcid.org/0009-0006-1020-1684
Kennichi Kakudo	https://orcid.org/0000-0002-0347-7264

Author Contributions

Conceptualization: ASH, MFH, KK. Methodology: ASH, AHA, DK, MM, KK. Software: AHA. Validation: ASH and MFH. Formal analysis: MM, DK. Investigation: ASH, KK. Resources: ASH. Data curation: AA, DPR, FQ, HA, HI, HC, IW, KM, MT, NMM, NK, NA, PA, YPS, ASH. Writing—original draft preparation: ASH, AHA, DK, MM. Writing—review and editing: AA, DPR, FQ, HA, HI, HC, IW, KM, MT, NMM, NK, NA, PA, YPS, ASH, KK. Visualization: ASH, AHA. Project administration: MM. Funding acquisition: ASH. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

K.K., a contributing editor of the *Journal of Pathology and Translational Medicine*, was not involved in the editorial evaluation or decision to publish this article. All remaining authors have declared no conflicts of interest.

Funding Statement

The authors would like to thank Universitas Indonesia for funding this research through a PUTI grant with contract number NKB-615/UN2.RST/HKP.05.00/2024.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024; 74: 229-63.
2. Baloch ZW, Mete O, Fadda G, et al. Papillary thyroid carcinoma. WHO classification of tumours series, 5th ed, Vol. 10. Endocrine and neuroendocrine tumours [Internet]. Lyon: International Agency for Research on Cancer, 2022 [cited 2024 Jul 15]. Available from: https://tumourclassification.iarc.who.int/chapters/53.
3. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in thyroid cancer incidence and mortality in the United States, 1974-2013. JAMA 2017; 317: 1338-48.
4. Wiltshire JJ, Drake TM, Uttley L, Balasubramanian SP. Systematic review of trends in the incidence rates of thyroid cancer. Thyroid 2016; 26: 1541-52.
5. Donnelly D, Geoghegan R, O'Brien C, Philbin E, Wheeler TS. Synthesis of heterocyclic-substituted chromones and related compounds as potential anticancer agents. J Med Chem 1965; 8: 872-5.
6. Liu Z, Bychkov A, Jung CK, et al. Interobserver and intraobserver variation in the morphological evaluation of noninvasive follicular thyroid neoplasm with papillary-like nuclear features in Asian practice. Pathol Int 2019; 69: 202-10.
7. Thompson LD, Poller DN, Kakudo K, Burchette R, Nikiforov YE, Seethala RR. An international interobserver variability reporting of the nuclear scoring criteria to diagnose noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a validation study. Endocr Pathol 2018; 29: 242-9.
8. Hirokawa M, Carney JA, Goellner JR, et al. Observer variation of encapsulated follicular lesions of the thyroid gland. Am J Surg Pathol 2002; 26: 1508-14.
9. 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: 736-44.
10. 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: 1336-40.
11. Zhu Y, Li Y, Jung CK, et al. Histopathologic assessment of capsular invasion in follicular thyroid neoplasms: an observer variation study. Endocr Pathol 2020; 31: 132-40.
12. Su HK, Wenig BM, Haser GC, et al. Inter-observer variation in the pathologic identification of minimal extrathyroidal extension in papillary thyroid carcinoma. Thyroid 2016; 26: 512-7.
13. 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: 1023-9.
14. Jung CK, Bychkov A, Song DE, et al. Molecular correlates and nuclear features of encapsulated follicular-patterned thyroid neoplasms. Endocrinol Metab (Seoul) 2021; 36: 123-33.
15. Harahap AS, Subekti I, Panigoro SS, et al. Profile of BRAFV600E, BRAFK601E, NRAS, HRAS, and KRAS mutational status, and clinicopathological characteristics of papillary thyroid carcinoma in Indonesian National Referral Hospital. Appl Clin Genet 2023; 16: 99-110.
16. Xin Y, Guan D, Meng K, Lv Z, Chen B. Diagnostic accuracy of CK-19, Galectin-3 and HBME-1 on papillary thyroid carcinoma: a meta-analysis. Int J Clin Exp Pathol 2017; 10: 8130-40.
17. Morariu EM, McCoy KL, Chiosea SI, et al. Clinicopathologic characteristics of thyroid nodules positive for the THADA-IGF2BP3 fusion on preoperative molecular analysis. Thyroid 2021; 31: 1212-8.
18. Chernock RD, Rivera B, Borrelli N, et al. Poorly differentiated thyroid carcinoma of childhood and adolescence: a distinct entity characterized by DICER1 mutations. Mod Pathol 2020; 33: 1264-74.
19. Seethala RR, Baloch ZW, Barletta JA, et al. Noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a review for pathologists. Mod Pathol 2018; 31: 39-55.
20. Farmer M, Petras RE, Hunt LE, Janosky JE, Galandiuk S. The importance of diagnostic accuracy in colonic inflammatory bowel disease. Am J Gastroenterol 2000; 95: 3184-8.
21. Kerkhof M, van Dekken H, Steyerberg EW, et al. Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. Histopathology 2007; 50: 920-7.
22. Xu B, Wang L, Tuttle RM, Ganly I, Ghossein R. Prognostic impact of extent of vascular invasion in low-grade encapsulated follicular cell-derived thyroid carcinomas: a clinicopathologic study of 276 cases. Hum Pathol 2015; 46: 1789-98.