ULTRA SONO GRAPHY

Assessing the diagnostic performance of thyroid biopsy with recommendations for appropriate interpretation

Su Min Ha[1,2], Jung Hwan Baek[3], Dong Gyu Na[4], Chan Kwon Jung[5], Chong Hyun Suh[3], Young Kee Shong[6], Tae-Yon Sung[7], Dong Eun Song[8], Jeong Hyun Lee[3]

[1]Department of Radiology and Thyroid Center, Chung-Ang University Hospital, Chung-Ang University College of Medicine, Seoul; [2]Department of Radiology and Research Institute of Radiology, Seoul National University Hospital, Seoul; [3]Department of Radiology and the Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul; [4]Department of Radiology, GangNeung Asan Hospital, Gangneung; [5]Department of Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul; Departments of [6]Internal Medicine, [7]Surgery, and [8]Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Purpose: The diagnostic performance of thyroid biopsy is influenced by several factors, including differences in the Bethesda categorization for malignancy, the inclusion or exclusion of non-diagnostic results, the definition used for the final diagnosis, and the definition of an inconclusive diagnosis. The purpose of this study was to provide an understanding of the factors influencing the diagnostic performance of thyroid biopsy.

Methods: We collected data retrospectively between January and December 2013 from a cohort of 6,762 thyroid nodules from 6,493 consecutive patients who underwent biopsy. In total, 4,822 nodules from 4,553 patients were included. We calculated the biopsy sensitivity according to the inclusion of different Bethesda categories in the numerator and the exclusion of non-diagnostic results, as well as the diagnostic accuracy according to different definitions of a benign diagnosis. We obtained the conclusive and inconclusive diagnosis rates.

Results: The sensitivity increased when more Bethesda categories were included in the numerator and when non-diagnostic results were excluded. When a benign thyroid nodule diagnosis was defined as benign findings on surgical resection, concordant benign results on at least two occasions, or an initial benign biopsy result and follow-up for more than 12 months, the accuracy was higher than when the diagnosis was based on surgical resection alone (91.1% vs. 68.7%). A higher conclusive diagnosis rate was obtained when Bethesda categories I and III were considered inconclusive than when Bethesda categories I, III and IV were considered inconclusive (78.3% vs. 72.8%, P<0.001).

Conclusion: Understanding the concepts presented herein is important in order to appropriately interpret the diagnostic performance of thyroid biopsy.

Keywords: Thyroid neoplasms; Thyroid nodule; Biopsy; Ultrasonography; Diagnosis

ORIGINAL ARTICLE

https://doi.org/10.14366/usg.19099
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2021;40:228-236

Received: December 31, 2019
Revised: May 19, 2020
Accepted: May 19, 2020

Correspondence to:
Jung Hwan Baek, MD, PhD,
Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea

Tel. +82-2-3010-4348
Fax. +82-2-476-0090
E-mail: radbaek@naver.com

How to cite this article:
Ha SM, Baek JH, Na DG, Jung CK, Suh CH, Shong YK, et al. Assessing the diagnostic performance of thyroid biopsy with recommendations for appropriate interpretation. Ultrasonography. 2021 Apr;40(2):228-236.

Introduction

Ultrasound (US)-guided biopsy is widely used to detect thyroid cancer, with satisfactory diagnostic performance [1–7]. Although many studies have evaluated the diagnostic performance of thyroid biopsy, including fine needle aspiration (FNA) and core needle biopsy (CNB), these studies have used heterogeneous definitions of benign or inconclusive biopsy results and there is a lack of consensus in the existing research on this topic. There have been no studies in which the investigators evaluated the fundamental factors affecting the interpretation of diagnostic performance. For example, a previous study [8] considering Bethesda category III as positive, rather than as an indeterminate result, added Bethesda category III to Bethesda categories IV, V, and VI in the numerator of the diagnostic performance calculation and the sensitivity of the biopsy marginally increased from 97.0% to 97.2%. Furthermore, studies differ in the definitions used for the final diagnosis. Two previous studies included surgical resection and clinical follow-up in the definition of the final diagnosis [9,10] whereas three previous studies [11–13] included only surgical resection. The impact of these unrealized factors is especially large in studies comparing the diagnostic performance of FNA and CNB. A recently published paper [14] found that there was no benefit in performing CNB over FNA and that both had a comparable diagnostic performance. However, that study [14] did not intentionally exclude non-diagnostic results of FNA after using propensity score-matching. Despite the overwhelming number of published studies, we suggest that there substantial variation exists in the interpretation of diagnostic performance across various studies and even within single studies.

The purpose of our study was to investigate the factors influencing the diagnostic performance of thyroid biopsy. Furthermore, we propose a recommendation for the appropriate interpretation of diagnostic performance.

Materials and Methods

Study Population

This retrospective study was approved by our institutional review board, and we received a waiver for informed written consent to use the data. The study population was obtained from 6,762 thyroid nodules from 6,493 consecutive patients who underwent biopsy between January and December 2013 at an academic tertiary referral hospital. Thyroid nodules in patients who had previously undergone biopsy (n=1,940), and 853 nodules without a final diagnosis were excluded. Finally, a total of 4,822 thyroid nodules with an initial biopsy from 4,553 patients were included in this study: 2,114 nodules from 1,928 patients who had undergone CNB and 2,708 nodules from 2,625 patients who had undergone FNA (Fig. 1). The study population has been analyzed in a previous study evaluating the efficacy and safety of CNB [15]. Whether to perform

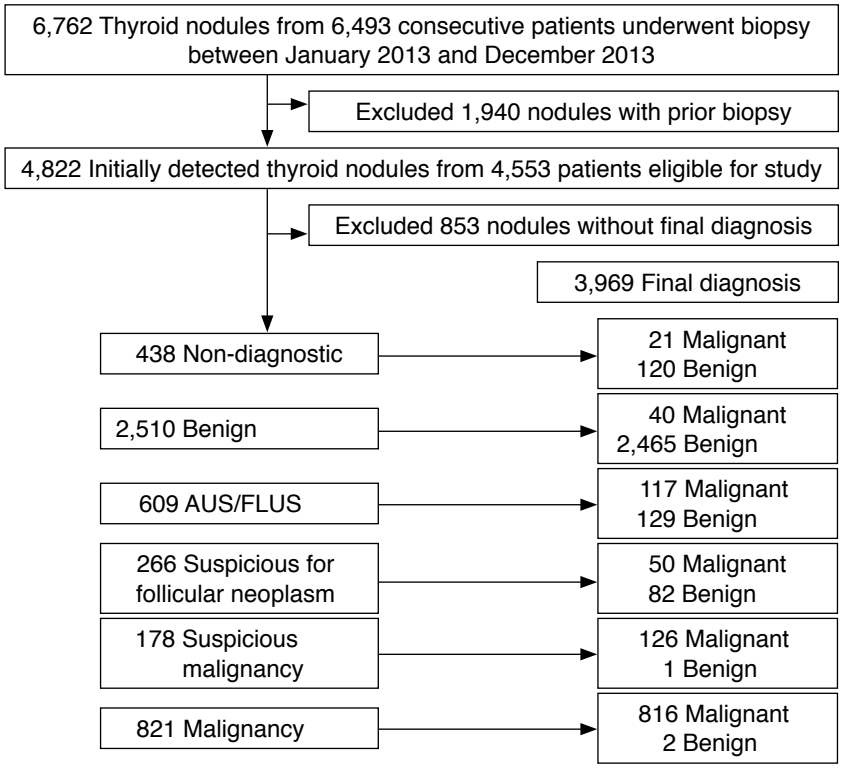

Fig. 1. **Patient flow and study outcomes in the study patients.** AUS/FLUS, atypia of undetermined significance/follicular lesion of undetermined significance.

CNB or FNA was determined mainly according to the referring physician's preference and CNB was performed for calcified nodules or predominantly cystic nodules, for which FNA may be less effective [16,17].

US-Guided FNA and CNB Procedures

US images were obtained for the evaluation of thyroid nodules using either an HDI 5000 (ATL Ultrasound, Bothell, WA, USA) or a Sequoia (Acuson, Mountain View, CA, USA) instrument equipped with a 5–12 MHz or an 8–15 MHz linear-array transducer. All US-guided procedures were performed by radiologists under the supervision of two faculty radiologists (J.H.B. and J.H.L., with 19 and 14 years of clinical experience, respectively, in performing and evaluating thyroid US). The US-guided CNB and FNA procedures for thyroid nodules were performed according to current practice guidelines [5,9,18–23].

Histopathologic Analysis of CNB Specimens and Cytopathologic Analysis of FNA

All CNB specimens and FNA cytological analyses were reviewed by a thyroid cytopathologist (D.E.S., with 11 years of clinical experience in thyroid cytopathology). Although the CNB diagnostic criteria for thyroid nodules had not yet been standardized during our study period, the histologic results of CNB were categorized into the same six categories of the Bethesda system that is used in the analysis of FNA cytology, with the following six standardized options [9,19,21,24,25]: Category I (non-diagnostic) included the absence of any identifiable follicular thyroid tissue, presence of only the normal thyroid gland, and tissue containing only a few follicular cells insufficient for diagnosis. Category II (benign) included all benign thyroidal and nonthyroidal disease. Category III (indeterminate lesion) corresponded to atypia of undetermined significance or follicular lesion of undetermined significance. Category IV (follicular neoplasm or suspicious for a follicular neoplasm) encompassed neoplastic lesions with follicular proliferative patterns. A category V (suspicious for malignancy) diagnosis was given when histologic features were strongly suspicious for malignancy, but insufficient for a definite diagnosis of malignancy. A category VI (malignancy) diagnosis was given when the typical histologic features were diagnosed as malignancy on a histologic specimen. The FNA cytology diagnoses were categorized into six categories according to the Bethesda System for Reporting Thyroid Cytopathology [9,24,26,27].

Analysis of US Findings

The US images were independently reviewed by two radiologists (J.H.B. and S.M.H.). When analyzing the US images, the radiologists assessed the thyroid nodules using criteria obtained from published reports [28–32], including the size (≥1 cm or <1 cm), internal content (solid, predominantly solid, predominantly cystic, or cystic), shape (round to oval or irregular), orientation (parallel or nonparallel), margin (well-defined smooth, microlobulated or spiculated, or ill-defined), echogenicity of the solid portion (hyperechogenicity or isoechogenicity, or hypoechogenicity or marked hypoechogenicity), and the presence of microcalcifications, macrocalcifications, and/or rim calcifications. The relationship between the final diagnosis (malignancy based on histopathologic findings from surgical resection or biopsy) and malignant US findings was assessed. The suspicious US features included irregular shape, nonparallel orientation, spiculated/microlobulated margin, marked hypoechogenicity, and the presence of microcalcifications [20].

Statistical Analysis

A final diagnosis of malignancy was made based on histopathologic readings from surgical resections or biopsies. A benign diagnosis was made when one of the following conditions was fulfilled: a surgical diagnosis of benignity, concordant benign results after biopsy on at least two occasions, or an initial benign biopsy result with a reduced or stable size on US follow-up at least 12 months later. We combined the diagnostic results of FNA and CNB thyroid biopsies. The diagnostic performance was calculated according to the following four criteria with different Bethesda categorizations in the numerator: criterion 1, Bethesda category VI; criterion 2, Bethesda categories V and VI; criterion 3, Bethesda categories IV, V, and VI; and criterion 4, Bethesda categories III, IV, V, and VI (Supplementary Table 1). We calculated the diagnostic performance including the diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We also calculated diagnostic performance parameters in the same manner after excluding Bethesda category I from the thyroid nodules in the dataset. We also analyzed diagnostic performance according to nodule size (<1 cm or ≥1 cm), different definitions of a benign diagnosis (i.e., surgical resection, concordant benign results on at least two occasions, or an initial benign biopsy result and follow-up at least 12 months later vs. surgical resection only), and differences in tumor subtype: (1) conventional papillary thyroid cancer (PTC) only; (2) conventional PTC and follicular variant PTC (FVPTC) and follicular carcinoma (FC); and (3) non-conventional PTC, including FC, FVPTC, medullary carcinoma, anaplastic carcinoma, sarcoma, lymphoma, and metastasis.

Nodules were classified and the inconclusive diagnosis rate was compared according to two criteria: criterion 1 (Bethesda categories I and III) and criterion 2 (Bethesda categories I, III, and IV). The

Student t-test was used for continuous variables. The chi-square test was used for comparisons of categorical variables. All tests were two-sided, and a significant P-value was defined as P<0.05. The statistical analysis was conducted using SPSS version 21.0 for Windows (IBM Corp., Armonk, NY, USA).

Results

The clinical and imaging characteristics of the nodules are shown in Table 1. In the FNA group, the mean size of the 2,708 nodules was 12.35±8.24 cm (range, 0.5 to 6.5 cm), with 54.4% (1,474 of 2,708) being ≥1 cm. In the CNB group, the mean size of the 2,114 nodules was 16.83±12.27 cm (range, 0.5–10 cm), with 70.0% (1,395 of 2,114) being ≥1 cm. The biopsy results and final diagnoses, according to the Bethesda category, are summarized in Table 2. The nodules (n=3,969) with a final diagnosis comprised 2,799 benign nodules and 1,170 malignant nodules (Supplementary Table 2).

Sensitivity for Malignancy

The final diagnoses of 3,969 thyroid nodules were included in the calculations of diagnostic performance (Table 3). The biopsy sensitivity was highest (94.8%) using criterion 4, and showed

Table 1. Clinical and ultrasonography characteristics of thyroid nodules

Characteristic	Total (n=4,822)	Final diagnosis (n=3,969)		P-value
		Benign (n=2,799)	Malignant (n=1,170)	
Age (year)	53.7±12.2	54.6±11.7	50.7±12.7	<0.001
Sex				
Male	1,086 (22.5)	549 (64.7)	300 (35.3)	<0.001
Female	3,736 (77.5)	2,250 (72.1)	870 (27.9)	
Nodule size (cm)	1.43±1.04	1.53±1.03	1.14±0.98	<0.001
<1	1,952 (40.5)	934 (57.3)	695 (42.7)	<0.001
≥1	2,870 (59.5)	1,865 (79.7)	475 (20.3)	
Composition				
Solid	3,588 (74.4)	1,895 (63.7)	1,081 (36.3)	<0.001
Partially solid	899 (18.6)	676 (89.5)	79 (10.5)	
Partially cystic	290 (6.0)	201 (96.6)	7 (3.4)	
Cystic	45 (1.0)	27 (90.0)	3 (10.0)	
Shape				
Ovoid to round	3,944 (81.8)	2,484 (77.7)	715 (22.4)	<0.001
Taller than wide	415 (8.6)	89 (23.8)	285 (76.2)	
Irregular	463 (9.6)	226 (57.1)	170 (42.9)	
Margin				
Smooth	2,756 (57.2)	1,781 (81.7)	400 (18.3)	<0.001
Spiculated	459 (9.5)	74 (17.3)	353 (82.7)	
Ill-defined	1,607 (33.3)	944 (69.4)	417 (30.6)	
Echogenicity				
Marked hypoechoic	540 (11.2)	146 (29.6)	348 (70.4)	<0.001
Hypoechoic	2,147 (44.5)	1,107 (63.0)	649 (37.0)	
Isoechoic	2,074 (43.0)	1,496 (89.6)	173 (10.4)	
Hyperechoic	61 (1.3)	50 (100)	0	
Calcifications				
Microcalcifications	441 (9.2)	126 (32.2)	265 (67.8)	<0.001
Macrocalcifications	656 (13.6)	294 (53.5)	256 (46.5)	
Rim calcifications	177 (3.7)	103 (72.5)	39 (27.5)	
None	3,548 (73.6)	2,276 (78.9)	610 (21.1)	

Values are presented as mean±standard deviation or number (%).

Table 2. Biopsy results according to the Bethesda category

Bethesda category	Final diagnosis (n=3,969)		Total (n=4,822)
	Benign (n=2,799)	Malignant (n=1,170)	
I	120 (4.3)	21 (1.8)	438 (9.1)
II	2,465 (88.1)	40 (3.4)	2,510 (52.1)
III	129 (4.6)	117 (10.0)	609 (12.6)
IV	82 (2.9)	50 (4.3)	266 (5.5)
V	1 (0.0)	126 (10.8)	178 (3.7)
VI	2 (0.0)	816 (69.7)	821 (17.0)

Values are presented as number (%).
I, non-diagnostic; II, benign; III, atypia of undetermined significance/follicular lesion of undetermined significance; IV, follicular neoplasm or suspicious for follicular neoplasm; V, suspicious for malignancy; VI, malignancy.

a steady increase moving from criterion 1 (69.7%) to criteria 2 (80.5%) and 3 (84.8%). With the exclusion of Bethesda category I (non-diagnostic) from the dataset, the sensitivity of all of the criteria increased: from 69.7% to 71.0% for criterion 1, from 80.5% to 82.0% for criterion 2, from 84.8% to 86.3% for criterion 3, and from 94.8% to 96.5% for criterion 4. When we assessed the diagnostic performance according to the nodule size, a higher sensitivity was obtained from biopsies of smaller nodules (<1 cm) than from biopsies of larger nodules (88.3% vs. 69.1% for criterion 2). Regarding tumor subtypes, a higher sensitivity and PPV were obtained when only conventional PTC was considered as a malignancy (Supplementary Table 3).

Diagnostic Accuracy According to Definitions of a Benign Diagnosis

We calculated the diagnostic performance according to different definitions of a benign diagnosis (Table 4). When a benign diagnosis was defined as a benign result upon surgical resection, concordant benign results on at least two occasions, or an initial benign biopsy result and follow-up of at least 12 months later-a definition that included more benign nodules, but with no change in the number of malignant nodules that were included-the specificity and NPV improved, with higher accuracy than was obtained when using the more strict definition of benign findings upon surgical resection only (91.1% to 68.7%). The sensitivity and PPV maintained similar rates regardless of the definition.

Conclusive Results

The conclusive diagnosis rate showed a significant difference (78.3% vs. 72.8%, P<0.001) when the inconclusive diagnosis rate was calculated using Bethesda categories I and III (21.7%, 1,047 of 4,822) and Bethesda categories I, III, and IV (27.2%, 1,313 of 4,822), respectively.

Discussion

Our study demonstrates several factors that may influence the diagnostic performance of thyroid nodule biopsy. The sensitivity increased when the numerator included more Bethesda categories and when nodules with non-diagnostic biopsy results were excluded from the dataset. Ideally, we recommend Bethesda categories V and VI or VI to be considered as positive for malignancy in the numerator, and that Bethesda category I nodules should not be excluded from the dataset for the interpretation of diagnostic performance. The diagnostic accuracy increased when a benign diagnosis was defined as benign findings on surgical resection, concordant benign results on at least two occasions, or an initial benign biopsy result and follow-up for more than 12 months. When conducting a diagnostic accuracy study, we suggest generating lower and higher bound estimates for accuracy by using surgical resection alone or by including other biopsy and follow-up data as the definition for the final diagnosis. The rate of conclusive results increased when we defined Bethesda categories I and III as inconclusive results compared to the combination of Bethesda categories I, III, and IV. In our opinion, the inconclusive rate may include Bethesda categories I and III, as they are candidates for diagnostic surgery or repeat biopsy. Our study results will be helpful in understanding the results of various diagnostic performance studies of thyroid biopsy.

The sensitivity is influenced by the datasets assigned to the numerator or denominator. First, regarding the numerator, there is heterogeneity in the application of the Bethesda system for thyroid biopsy interpretation. In a previous study [8] that considered Bethesda category III as positive, rather than as an indeterminate result, adding Bethesda category III to Bethesda categories IV, V, and VI in the numerator marginally increased the sensitivity of thyroid FNA from 97.0% to 97.2%. Regarding the denominator, in a recent study by Choi et al. [14] comparing the diagnostic performance of thyroid biopsy procedures for detecting malignancy, excluding Bethesda category IV from the denominator increased the sensitivity of FNA and CNB sensitivity from 93.8% to 94.0% and from 84.7% to 88.1%, respectively, and excluding Bethesda categories I, III, and IV from the denominator substantively increased the sensitivity even further, to 99.8% and 99.1%, respectively. Accordingly, an unrealistically high sensitivity will be calculated when malignancies classified as Bethesda categories I, III, and IV are excluded from the denominator due to the reduced numbers of false negative results. We also observed sensitivity changes with the exclusion of Bethesda category I from the dataset. Therefore, if a study excludes the majority of non-diagnostic biopsy results from the analysis, the sensitivity will be biased, especially when comparing FNA and CNB, which have significantly different non-diagnostic result rates.

Table 3. Changes in the sensitivity for malignancy according to the criteria for a positive diagnosis

	Criterion 1	Criterion 2	Criterion 3	Criterion 4	P-value[a]
Total (n=3,969)					
TP	816	942	992	1,109	
TN	2,797	2,796	2,714	2,585	
FP	2	3	85	214	
FN	354	228	178	61	
Sensitivity	69.7	80.5	84.8	94.8	<0.001
Specificity	99.9	99.9	97.0	92.4	<0.001
Accuracy	91.0	94.2	93.4	93.1	0.004
PPV	99.8	99.7	92.1	83.8	
NPV	88.8	92.5	93.8	97.7	
Total[b] (n=3,828)					
TP	816	942	992	1,109	
TN	2,677	2,676	2,594	2,465	
FP	2	3	85	214	
FN	333	207	157	40	
Sensitivity	71.0	82.0	86.3	96.5	<0.001
Specificity	99.9	99.9	96.8	92.0	<0.001
Accuracy	91.2	94.5	93.7	93.4	0.004
PPV	99.8	99.7	92.1	83.8	
NPV	88.9	92.8	94.3	98.4	
Size <1 cm (n=1,629)					
TP	518	614	619	665	
TN	933	932	929	899	
FP	1	2	5	35	
FN	177	81	76	30	
Sensitivity	74.5	88.3	89.1	95.7	<0.001
Specificity	99.9	99.8	99.5	96.3	<0.001
Accuracy	89.1	94.9	95.0	96.0	<0.001
PPV	99.8	99.7	99.2	95.0	
NPV	84.1	84.1	92.4	96.8	
Size ≥1 cm (n=2,340)					
TP	298	328	373	444	
TN	1,864	1,864	1,785	1,686	
FP	1	1	80	179	
FN	177	147	102	31	
Sensitivity	62.7	69.1	78.5	93.5	<0.001
Specificity	99.9	99.9	95.7	90.4	<0.001
Accuracy	92.4	93.7	92.2	91.0	0.004
PPV	99.7	99.7	82.3	71.3	
NPV	91.3	92.7	94.6	98.2	

Criterion 1, Bethesda category VI (malignancy); Criterion 2, Bethesda category VI (malignancy) and Bethesda category V (suspicious for malignancy); Criterion 3, Bethesda category VI (malignancy), Bethesda category V (suspicious for malignancy), and Bethesda category IV (follicular neoplasm or suspicious for a follicular neoplasm); Criterion 4, Bethesda category VI (malignancy), Bethesda category V (suspicious for malignancy), Bethesda category IV (follicular neoplasm or suspicious for a follicular neoplasm), and Bethesda category III (atypia of undetermined significance/follicular lesion of undetermined significance).

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

[a]P-values for trends of sensitivity, specificity, and accuracy were calculated using generalized estimating equations. [b]Exclusion of Bethesda category I (non-diagnostic) from the dataset.

Table 4. Changes in diagnostic accuracy according to different definitions of benign thyroid nodule diagnoses

	Surgical resection (n=1,055)	Surgical resection or concordant benign diagnosis	
		At least two occasions (n=1,200)	At least two occasions or initially benign and 1-year follow-up (n=3,969)
TP	590	590	590
TN	135	279	2,797
FP	1	2	2
FN	329	329	329
Sensitivity	64.2	64.2	64.2
Specificity	99.3	99.3	99.9
Accuracy	68.7	72.4	91.1
PPV	99.8	99.7	99.7
NPV	29.1	45.9	89.5

Diagnostic performance was calculated according to criterion 1, which included Bethesda category VI (malignancy).
TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; NPV, negative predictive value.

Based on our analysis, including Bethesda categories V and VI or VI as positive for malignancy in the numerator and not excluding Bethesda category I from the dataset appear to be the most recommended conditions for diagnostic interpretation.

As the definition used for the final diagnosis can affect the results of diagnostic accuracy, an appropriate definition of the final diagnosis is critical. Two previous studies including surgical resection and clinical follow-up as the definition of the final diagnosis [9,10] showed higher accuracy than three studies [11–13] that defined the final diagnosis on the basis of surgical resection alone. Our results verified that when the definition of the final diagnosis was broader, with the inclusion of more benign thyroid nodules, the specificity and ensuing accuracy were higher than when a stricter definition was used, such as surgical resection alone. The sensitivity and PPV were maintained due to the absence of a change in the number of malignant nodules diagnosed in these different scenarios. Therefore, we suggest generating a range of lower and upper bound estimates of diagnostic accuracy corresponding to the use of surgical resection alone or surgical resection combined with biopsy and follow-up data as the definitions of the final diagnosis.

Several studies have applied the terms "conclusive" and "inconclusive" when comparing results across different biopsy procedures of thyroid nodules. A higher rate of conclusive biopsy results is favorable, as further unnecessary biopsies can then be minimized [29,33]. However, the definition of these two terms is inconsistent. For example, in one study [34], inconclusive results included Bethesda categories I, III, and IV, whereas other studies [20,35] defined inconclusive results as including Bethesda categories I and III. If we simulate the previous findings of Suh et al. [20] by classifying only Bethesda category IV as an inconclusive result, the conclusive result rate increases from 5.9% (Bethesda categories I and III) to 9.2%. The conclusive and inconclusive rates are inversely proportional, indicating that as one increases, the other decreases. The 2017 Bethesda system considers Bethesda categories IV, V, and VI to be conclusive results [26]. Therefore, we suggest that the most appropriate definition of inconclusive results would be Bethesda category I and III nodules.

Our study revealed other possible factors influencing the diagnostic performance of thyroid biopsy. The diagnostic performance may be influenced when conventional PTC prevails or the proportion of PTC among malignant tumors is relatively high in the patient population [36]. Regarding noninvasive follicular thyroid neoplasms with papillary-like nuclear features, which are included in the revised Bethesda system [26], if this diagnosis frequently occurs, we hypothesize that the diagnostic performance of procedures would be underestimated, although we did not specifically evaluate this possibility. Regarding the possibility of nodule size as another possible contributor to bias, as FC and FVPTC are usually larger than conventional PTC, better sensitivity was observed in smaller nodules. This result is similar to that of a previous study concerning CNB, which found higher sensitivity in small nodules, with a greater proportion of conventional PTC [20]. Therefore, when we interpret the diagnostic performance, we should consider the proportion of conventional PTCs and the tumor size in the cohort.

As another important factor affecting the diagnostic performance of thyroid biopsy, the proportion of repeated biopsies of nodules with previous inconclusive diagnostic results should be considered and matched in the patient population in order to obtain the optimal comparison of thyroid biopsy procedures using different patient populations. The repeated biopsy of nodules with prior inconclusive results generally yields a higher rate of repeated inconclusive results and a lower diagnostic sensitivity for malignancy compared to the initial biopsy results [7,37], which may cause a biased comparison if it is not matched between two populations.

The major limitation of our study is that there may have been selection bias due to its retrospective study design, and there may have been inherent bias in terms of the patient selection. Our study should be interpreted with some reservations because of the possibility of selection bias towards suspicious nodules owing to the usage of US for the CNB group, and because the biopsy procedure was determined according to the referring physician's preference. However, our large study population may compensate for this selection bias. As mentioned above, the proportion of repeated biopsies of nodules with previous inconclusive diagnostic results

should be considered, which we did not investigate. Future research with a high-volume dataset including either FNA- or CNB-diagnosed thyroid nodules would be beneficial to minimize these limitations. In addition, this study was carried out at a single institution, and therefore further generalization is required in future, multi-center studies. Lastly, most of the benign nodules were not confirmed with surgery.

In conclusion, this study suggests some factors that may influence the diagnostic performance of thyroid biopsy. Understanding these concepts is important for a more critical and appropriate interpretation of diagnostic performance.

ORCID: Su Min Ha: https://orcid.org/0000-0002-1833-0919; Jung Hwan Baek: https://orcid.org/0000-0003-0480-4754; Dong Gyu Na: https://orcid.org/0000-0001-6422-1652; Chan Kwon Jung: https://orcid.org/0000-0001-6843-3708; Chong Hyun Suh: https://orcid.org/0000-0002-4737-0530; Young Kee Shong: https://orcid.org/0000-0002-7911-9471; Tae-Yon Sung: https://orcid.org/0000-0002-2179-6269; Dong Eun Song: https://orcid.org/0000-0002-9583-9794; Jeong Hyun Lee: https://orcid.org/0000-0002-0021-4477

Author Contributions

Conceptualization: Baek JH, Ha SM. Data acquisition: Baek JH, Ha SM, Suh CH, Shong YK, Sung TY, Song DE. Data analysis or interpretation: Ha SM. Drafting of the manuscript: Baek JH, Jung CK, Lee JH. Critical revision of the manuscript: Baek JH, Na DG. Approval of the final version of the manuscript: all authors.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Supplementary Material

Supplementary Table 1. Criteria for test positive were defined according to the four criteria (https://doi.org/10.14366/usg.19099).

Supplementary Table 2. Histopathologic results of 1,170 malignancies (https://doi.org/10.14366/usg.19099).

Supplementary Table 3. Overestimation of the sensitivity according to the carcinoma subtype (https://doi.org/10.14366/usg.19099).

References

1. Goudy SL, Flynn MB. Diagnostic accuracy of palpation-guided and image-guided fine-needle aspiration biopsy of the thyroid. Ear Nose Throat J 2005;84:371-374.
2. Cesur M, Corapcioglu D, Bulut S, Gursoy A, Yilmaz AE, Erdogan N, et al. Comparison of palpation-guided fine-needle aspiration biopsy to ultrasound-guided fine-needle aspiration biopsy in the evaluation of thyroid nodules. Thyroid 2006;16:555-561.
3. Cai XJ, Valiyaparambath N, Nixon P, Waghorn A, Giles T, Helliwell T. Ultrasound-guided fine needle aspiration cytology in the diagnosis and management of thyroid nodules. Cytopathology 2006;17:251-256.
4. Baek JH. Current status of core needle biopsy of the thyroid. Ultrasonography 2017;36:83-85.
5. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14:587-595.
6. Son HM, Kim JH, Kim SC, Yoo RE, Bae JM, Seo H, et al. Distribution and malignancy risk of six categories of the pathology reporting system for thyroid core-needle biopsy in 1,216 consecutive thyroid nodules. Ultrasonography 2020;39:159-165.
7. Hong MJ, Na DG, Kim SJ, Kim DS. Role of core needle biopsy as a first-line diagnostic tool for thyroid nodules: a retrospective cohort study. Ultrasonography 2018;37:244-253.
8. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda System for Reporting Thyroid Cytopathology: a meta-analysis. Acta Cytol 2012;56:333-339.
9. Na DG, Kim JH, Sung JY, Baek JH, Jung KC, Lee H, et al. Core-needle biopsy is more useful than repeat fine-needle aspiration in thyroid nodules read as nondiagnostic or atypia of undetermined significance by the Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2012;22:468-475.
10. Sung JY, Na DG, Kim KS, Yoo H, Lee H, Kim JH, et al. Diagnostic accuracy of fine-needle aspiration versus core-needle biopsy for the diagnosis of thyroid malignancy in a clinical cohort. Eur Radiol 2012;22:1564-1572.
11. Hakala T, Kholova I, Sand J, Saaristo R, Kellokumpu-Lehtinen P. A core needle biopsy provides more malignancy-specific results than fine-needle aspiration biopsy in thyroid nodules suspicious for malignancy. J Clin Pathol 2013;66:1046-1050.
12. Karstrup S, Balslev E, Juul N, Eskildsen PC, Baumbach L. US-guided fine needle aspiration versus coarse needle biopsy of thyroid nodules. Eur J Ultrasound 2001;13:1-5.
13. Renshaw AA, Pinnar N. Comparison of thyroid fine-needle aspiration and core needle biopsy. Am J Clin Pathol 2007;128:370-374.
14. Choi YJ, Baek JH, Park HS, Shim WH, Kim TY, Shong YK, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. Thyroid 2017;27:546-552.
15. Suh CH, Baek JH, Choi YJ, Kim TY, Sung TY, Song DE, et al. Efficacy and safety of core-needle biopsy in initially detected thyroid nodules via propensity score analysis. Sci Rep 2017;7:8242.
16. Ha EJ, Baek JH, Lee JH, Song DE, Kim JK, Shong YK, et al.

Sonographically suspicious thyroid nodules with initially benign cytologic results: the role of a core needle biopsy. Thyroid 2013;23:703-708.
17. Ha EJ, Baek JH, Lee JH, Kim JK, Kim JK, Lim HK, et al. Core needle biopsy can minimise the non-diagnostic results and need for diagnostic surgery in patients with calcified thyroid nodules. Eur Radiol 2014;24:1403-1409.
18. Bae SY, Kim S, Lee JH, Lee HC, Lee SK, Kil WH, et al. Poor prognosis of single hormone receptor- positive breast cancer: similar outcome as triple-negative breast cancer. BMC Cancer 2015;15:138.
19. Choi YJ, Baek JH, Ha EJ, Lim HK, Lee JH, Kim JK, et al. Differences in risk of malignancy and management recommendations in subcategories of thyroid nodules with atypia of undetermined significance or follicular lesion of undetermined significance: the role of ultrasound-guided core-needle biopsy. Thyroid 2014;24:494-501.
20. Suh CH, Baek JH, Lee JH, Choi YJ, Kim JK, Sung TY, et al. The role of core-needle biopsy as a first-line diagnostic tool for initially detected thyroid nodules. Thyroid 2016;26:395-403.
21. Yeon JS, Baek JH, Lim HK, Ha EJ, Kim JK, Song DE, et al. Thyroid nodules with initially nondiagnostic cytologic results: the role of core-needle biopsy. Radiology 2013;268:274-280.
22. Na DG, Baek JH, Jung SL, Kim JH, Sung JY, Kim KS, et al. Core needle biopsy of the thyroid: 2016 consensus statement and recommendations from Korean Society of Thyroid Radiology. Korean J Radiol 2017;18:217-237.
23. Chung SR, Baek JH, Choi YJ, Sung TY, Song DE, Kim TY, et al. The role of core needle biopsy for the evaluation of thyroid nodules with suspicious ultrasound features. Korean J Radiol 2019;20:158-165.
24. Cibas ES, Ali SZ; NCI Thyroid FNA State of the Science Conference. The Bethesda System For Reporting Thyroid Cytopathology. Am J Clin Pathol 2009;132:658-665.
25. Jung CK, Baek JH. Recent advances in core needle biopsy for thyroid nodules. Endocrinol Metab (Seoul) 2017;32:407-412.
26. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2017;27:1341-1346.
27. Jung CK, Min HS, Park HJ, Song DE, Kim JH, Park SY, et al. Pathology reporting of thyroid core needle biopsy: a proposal of the Korean Endocrine Pathology Thyroid Core Needle Biopsy Study Group. J Pathol Transl Med 2015;49:288-299.
28. Frates MC, Benson CB, Charboneau JW, Cibas ES, Clark OH, Coleman BG, et al. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. Ultrasound Q 2006;22:231-238.
29. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1-133.
30. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. Radiology 2011;260:892-899.
31. Moon WJ, Jung SL, Lee JH, Na DG, Baek JH, Lee YH, et al. Benign and malignant thyroid nodules: US differentiation: multicenter retrospective study. Radiology 2008;247:762-770.
32. Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology consensus statement and recommendations. Korean J Radiol 2016;17:370-395.
33. Trimboli P, Giovanella L. Reliability of core needle biopsy as a second-line procedure in thyroid nodules with an indeterminate fine-needle aspiration report: a systematic review and meta-analysis. Ultrasonography 2018;37:121-128.
34. Zhang M, Zhang Y, Fu S, Lv F, Tang J. Thyroid nodules with suspicious ultrasound findings: the role of ultrasound-guided core needle biopsy. Clin Imaging 2014;38:434-438.
35. Choi SH, Baek JH, Lee JH, Choi YJ, Hong MJ, Song DE, et al. Thyroid nodules with initially non-diagnostic, fine-needle aspiration results: comparison of core-needle biopsy and repeated fine-needle aspiration. Eur Radiol 2014;24:2819-2826.
36. VanderLaan PA, Marqusee E, Krane JF. Features associated with locoregional spread of papillary carcinoma correlate with diagnostic category in the Bethesda System for Reporting Thyroid Cytopathology. Cancer Cytopathol 2012;120:245-253.
37. Na DG, Min HS, Lee H, Won JK, Seo HB, Kim JH. Role of core needle biopsy in the management of atypia/follicular lesion of undetermined significance thyroid nodules: comparison with repeat fine-needle aspiration in subcategory nodules. Eur Thyroid J 2015;4:189-196.