narra j

Review Article

Diagnostic accuracy of preoperative ultrasonography-guided fine-needle aspiration biopsy in distinguishing malignancy in large thyroid nodules: A systematic review, meta-analysis, and meta-regression

Putri O. Zulfa[1], Muhammad Iqhrammullah[2] and Hendra Zufry[3,4]*

[1]Medical Research Unit, Faculty of Medicine, Universitas Syiah Kuala, Banda Aceh, Indonesia; [2]Faculty of Public Health, Universitas Muhammadiyah Aceh, Banda Aceh, Indonesia; [3]Division of Endocrinology, Metabolic, and Diabetes, Department of Internal Medicine, Faculty of Medicine, Universitas Syiah Kuala, Banda Aceh, Indonesia; [4]Division of Endocrinology, Metabolic, and Diabetes, Department of Internal Medicine, Dr. Zainoel Abidin Hospital, Banda Aceh, Indonesia

*Corresponding author: hendra_zufry@usk.ac.id

Abstract

Controversy persists regarding the effectiveness of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancies in large thyroid nodules. The prevailing belief that larger thyroid nodules inherently pose a higher risk of malignancy has led to a common practice of suggesting thyroidectomy for large thyroid nodules. Herein, the aim of this study was to assess the diagnostic accuracy of preoperative US-FNAB for distinguishing malignancy in large thyroid nodules. A search for published records was carried out on October 20, 2023, utilizing the search feature available on PubMed, Scopus, Embase, and Google Scholar. Patients with large thyroid nodules (3 cm or larger) who underwent preoperative US-FNAB and postoperative histopathological tests were included. Related outcomes, including false positive, false negative, true negative, true positive, specificity, and sensitivity, were extracted from each study. Pooled specificity and sensitivity were estimated, and the summarized receiver operating characteristic (sROC) curve, along with the summarized area under the curve (sAUC), was calculated. Out of 133 articles identified across four databases, ten studies with a total sample of 2752 patients were included. The overall diagnostic sensitivity was 72% (95%CI: 50–86%; p=0.00) and specificity was 96% (95%CI: 87–90%; p=0.00). The positive predictive value (PPV) and negative predictive value (NPV) were 93% (95%CI: 89–98%) and 75% (95%CI: 72–79%), respectively. sAUC was 93%, suggesting the diagnostic tool is accurate. Meta-regression analysis revealed that factors such as the number of samples, country (high-income vs upper-middle income), demographic characteristics (age and sex), and different thyroid size cut-off values did not significantly impact the sensitivity or specificity of US-FNAB. In conclusion, the present study confirms the reliability of US-FNAB in distinguishing malignancy in large thyroid nodules, emphasizing its role in reducing unnecessary thyroidectomy by identifying high-risk patients and challenging the conventional practice of routine thyroidectomy for large thyroid nodules.

Keywords: Large thyroid nodule, fine-needle aspiration, accuracy, histopathology, review

Introduction

Thyroid nodules are prevalent endocrine pathologies that often require surgical intervention. Additionally, it has seen a substantial increase in malignancy rates globally [1], ranging from 5%

https://narraj.org/

Received: July 5, 2024 - Accepted: January 9, 2025 - Published online: February 25, 2025

to 20% [2]. Globally, in 2022, the diagnosed cases of thyroid nodules reached 9,276,178 individuals, with an overall prevalence of 24.83% regardless of the diagnostic techniques [3]. Initial assessment of a thyroid nodule involves evaluating thyroid function, identifying clinical risk factors, and performing neck imaging studies [4]. Ultrasound is considered the gold standard for evaluating the morphology of thyroid nodules [5], while biopsy remains the definitive method for determining whether a thyroid nodule is benign or malignant [6]. However, the contribution of overdiagnosis to the increasing incidence of thyroid malignancy was substantial [7]. Despite various diagnostic techniques, the challenge lies in accurately distinguishing between benign and malignant nodules; while thyroid function tests, scintigraphy, and ultrasonography provide essential diagnostic information, they fall short in precise differentiation [8].

Thyroid nodules exceeding 3 cm are frequently managed with total thyroidectomy or lobectomy due to an elevated risk of malignancy despite benign findings on ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) [9,10]. US-FNAB is widely acknowledged as the primary diagnostic tool for thyroid nodules due to its simplicity, safety, and cost-effectiveness [11-13]. The ability to accurately stratify patients with thyroid nodules preoperatively is imperative because most thyroid nodules are benign [14]. However, questions have arisen about its reliability, particularly in large thyroid nodules (3 cm or larger), leading to persistent controversy regarding its effectiveness in distinguishing malignancy [15].

Previous studies have shown that malignancy risk increases for thyroid nodules up to 2 cm but plateaus beyond this size, with larger nodules generally exhibiting lower malignancy rates [16,17]. However, the prevailing belief that larger thyroid nodules inherently pose a higher malignancy risk has led to a common practice of suggesting thyroidectomy for large thyroid nodules [18]. Thyroidectomy, a procedure to remove all or part of the thyroid gland, is commonly performed for thyroid nodules yet carries significant risks such as hypocalcemia, injury to the recurrent laryngeal nerve, and the need for lifelong thyroid hormone replacement therapy [19]. To the best of the author's knowledge, no systematic review and meta-analysis have compared the diagnostic accuracy of preoperative US-FNAB with post-histopathological testing for large thyroid nodules. Herein, the aim of this study was to assess the diagnostic accuracy of preoperative US-FNAB compared to postoperative histopathology tests for distinguishing malignancy in large thyroid nodules.

Methods

Study design and setting

A systematic review, meta-analysis, and meta-regression were conducted. Protocols for the present systematic review and meta-analysis were designed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [20]. The research question was focused on determining the sensitivity, specificity, positive likelihood value, negative likelihood value, and diagnostic odds ratio of preoperative US-FNAB compared to postoperative histopathology tests in distinguishing malignancy in large thyroid nodules.

Inclusion and exclusion criteria

Inclusion and exclusion criteria in the present study adopted the PICOS framework (Population, Intervention, Control, Outcome, and Study design). The population consists of patients with large thyroid nodules (≥3 cm) undergoing diagnostic evaluation. Intervention, or index test, was preoperative US-FNAB, while the comparator was the postoperative histopathological examination, considered the gold standard. Outcomes of interest focus on diagnostic accuracy measures, including true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), sensitivity, and specificity. Regarding the study design, an observational study (e.g., cross-sectional, cohort studies) was included. No publication year restriction was applied, the search included studies from database inception to October 20, 2023. Studies were excluded if having one of the following criteria: (1) published in a language other than English and Indonesia; (2) not reporting one of the outcomes of interest; (3) review articles, conference abstracts, case reports, case series, editorials, commentary, thesis, and erratum.

Search strategy

Search for the published records was carried out on October 20, 2023 utilizing the searching feature available on PubMed, Scopus, Embase, and Google Scholar. Boolean operators ('OR'/'AND') were employed in the four databases. The keywords 'thyroid,' 'thyroid nodule,' 'fine-needle aspiration,' 'fnab,' and 'large' were used (**Table 1**). An advanced search was employed, with the filter set to include only titles.

Table 1. Combined keywords using Boolean operations employed in each database

Database	Keywords
Embase	(('thyroid nodule' OR thyroid) AND ('fine-needle' OR fnab) AND (large OR larger))
Google Scholar	TITLE (('thyroid nodule' OR thyroid) AND ('fine-needle' OR fnab) AND (large OR larger))
PubMed	(('thyroid nodule'[Title] OR thyroid [Title]) AND ('fine-needle'[Title] OR fnab[Title]) AND (large[Title] OR larger[Title]))
Scopus	TITLE (('thyroid nodule' OR thyroid) AND ('fine-needle' OR fnab) AND (large OR larger))

Screening and selection of the records

PRISMA was employed to guide our screening and selection process, which was carried out by two independent reviewers (POZ and MI). Duplicates were immediately removed once the identified records were imported to Zotero v.6.0.30 (https://www.zotero.org/). The screening process was conducted based on the 'Title' and 'Abstract,' utilizing ASReview (https://asreview.nl/), an open-source software designed for efficient systematic review screening through machine learning-assisted prioritization [21]. ASReview enabled the rapid and accurate identification of potentially relevant studies by ranking titles and abstracts according to their likelihood of meeting the eligibility criteria [21]. The screening process was then manually conducted by two independent reviewers (POZ and MI) who further refined the selection through full-text screening based on the predetermined inclusion and exclusion criteria. Any discrepancies were resolved through consultation with the supervisor (HZ).

Data extraction

Two independent reviewers (POZ and MI) were involved in the data extraction, with any discrepancies resolved through consultation with the supervisor (HZ). From the included studies, study characteristics, including author name, publication year, country, study design, and sample size. Patients' characteristics encompassed mean age, total male and female patients, and cut-off values (>3 cm, ≥4 cm, or >4 cm) for thyroid nodules. Related outcomes, such as actual positive, actual negative, FP, FN, TN, TP, specificity, and sensitivity, were extracted from each study. The extracted outcomes were presented as frequencies (e.g., actual positives, actual negatives, TP, FP, TN, FN) and as decimals (e.g., specificity, sensitivity). In instances where FP, FN, TN, and TP were unreported, the data were approximated from specificity and sensitivity values. Actual positive refers to cases classified as malignant based on the Bethesda System for Reporting Thyroid Cytopathology [22], while actual negative denotes benign cases according to the same classification criteria [22]. When data for actual positives were absent, the number of positives was estimated using the formula: (1 - specificity) × sample size. In cases where actual negatives were absent or unavailable, the formula (1 - sensitivity) × sample size was employed. For TN, the estimation was derived by multiplying the specificity by the number of actual negatives. Similarly, FN was calculated using the formula: (1 - sensitivity) × actual negative, and FP was determined as (1 - specificity) × actual positive. Finally, TP was estimated by multiplying sensitivity with the number of actual positives.

Quality assessment

Quality assessment of the included studies was conducted by two independent reviewers (POZ and MI), with any discrepancies resolved through consultation with the supervisor (HZ). Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2, a validated framework designed to evaluate potential sources of bias and applicability concerns in diagnostic accuracy studies, was employed to assess the quality of included studies [23]. This assessment tool comprises four domains: 'patient selection,' 'index test,' 'reference standard,' and 'flow and timing.' Each domain

contains signaling questions (yes/no/unclear) designed to identify potential risks of bias or applicability concerns. Based on these answers, a judgment of 'low risk,' 'some concerns,' or 'high risk' is made for each domain. The results of the QUADAS-2 assessment were summarized in a bar ranging from 0 to 100%, created using Microsoft Excel v.2021 (Microsoft Inc., Washington, USA), illustrating the risk of bias and applicability concerns across all domains for the included studies.

Meta-analysis and meta-regression

Diagnostic meta-analysis utilized the 'midas' package in STATA v.17 (StataCorp LLC, Texas, USA) [24]. A two-level mixed-effect logistic regression model with independent binomial distribution was employed. Heterogeneity was determined by I^2<50% and p<0.1, employing a bivariate random-effect model in cases of heterogeneity. A goodness-of-fit analysis was conducted to assess the appropriateness of the meta-analysis model; Mahalanobis D-squared statistic identified potential outliers, while Chi-squared quantile tested how well the data fit the theoretical distribution. Bivariate normality test using deviance residuals detected discrepancies between observed and expected values, supporting the normal distribution assumption and confirming the model's accuracy.

Pooled specificity and sensitivity were estimated, and the summarized receiver operating characteristic (sROC) curve, along with the summarized area under the curve (sAUC), was calculated. sAUC was used to evaluate diagnostic accuracy, with ≥0.75 indicating good accuracy and ≥0.90 reflecting excellent accuracy. An AUC of ≥0.75 was set as the criterion for acceptable model performance in the present study. Cook's distance and outlier detection plots were employed to assess the influence of individual studies on the pooled estimate, while Deeks' funnel plot was used to evaluate publication bias. The restricted maximum likelihood method was used for meta-regression of the subgroup analysis.

Sub-group analysis was conducted based on the total sample, sex, age, thyroid size cut-off values, and country. The number of samples and age were treated as continuous variables. The male-to-female ratio was categorized based on the dominance of either sex. Two binary variables were used for male representation: "Male>50%" for studies with more than 50% male patients and "Male<50%" for studies with fewer than 50%. Thyroid size cut-offs were categorized into ≥4 cm and ≥3 cm. The country classification was based on economic status, with studies grouped into high-income and upper-middle-income countries, according to the World Bank.

Results

Characteristics of the included studies

The literature search and selection process workflow and the number of publications obtained from each step are presented in **Figure 1**. The present study identified 133 articles through a systematic search across four databases. After removing duplicates, 30 out of 55 studies were excluded based on the title and abstract screening using predefined inclusion and exclusion criteria. During the full-text review, 12 studies were excluded for being irrelevant to the research objectives, 1 was a conference abstract, 1 was an editorial, and 1 was a case report. Finally, 10 studies, with a total sample of 2752 patients, were included in the present systematic review (**Table 2**).

The included studies were conducted in various countries: Turkey (n=3), the United States of America (n=2), South Korea (n=3), France (n=1), and Indonesia (n=1). The mean age ranges from 44.4 to 53 years, with four studies not reporting the mean age (N/A) [13,25-27]. Two studies did not provide the standard deviation (SD) but still reported the mean age (50 and 47.8 years) [28,29]. Female participants were more prevalent in most studies, with some studies reporting a higher number of females than males (e.g., 117 males vs 544 females, 80 males vs 183 females). In contrast, other studies have more males than females (e.g., 79 males vs 11 females). Two studies did not report sex data (N/A) [13,30]. The cut-off values of thyroid nodule size included in each study started from ≥3 cm (n=4), ≥4 cm (n=3), and >4 cm (n=3). The sensitivity of US-FNAB compared to the postoperative histopathological test ranged from 0.80 to 1, whereas specificity ranged from 0.15 to 0.97 (**Table 2**).

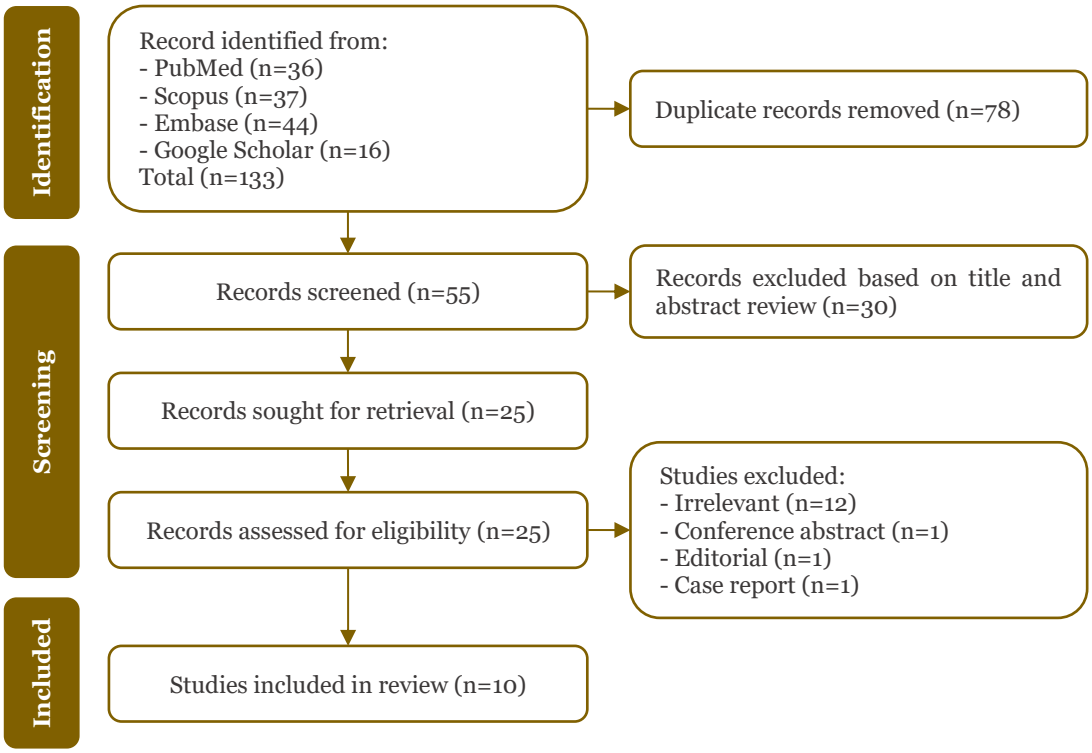

Figure 1. PRISMA flowchart of the included studies.

Quality assessment

Most of the included studies indicated a low risk of bias (90%) (**Figure 2**). However, in the study conducted by Yoon *et al.*, the risk of bias was deemed some concerns about patient selection due to unspecified criteria and uncertain numbers of patients who underwent US-FNAB also underwent thyroidectomy [25].

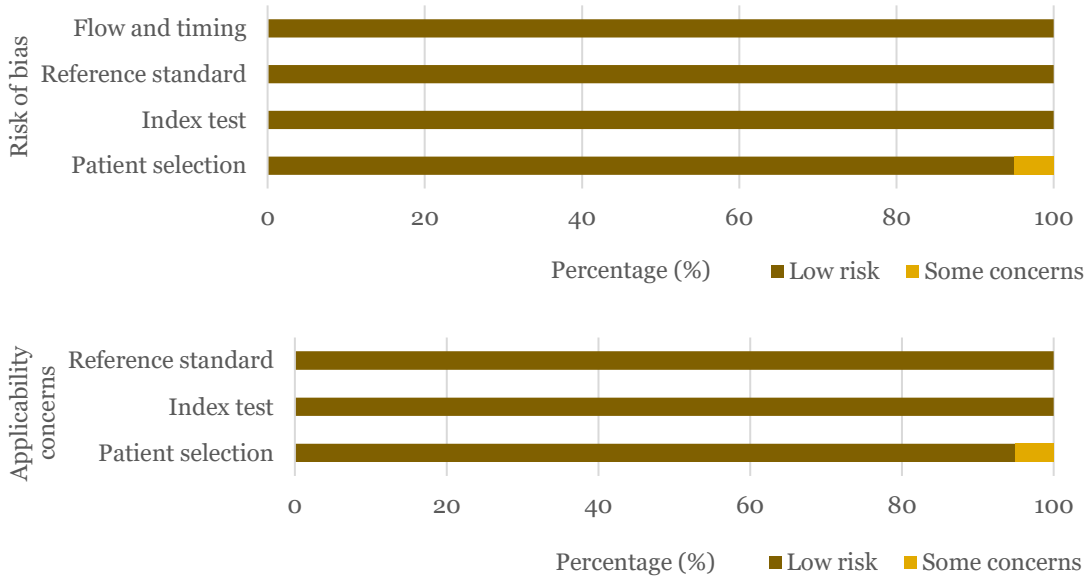

Figure 2. Quality assessment of the included studies based on the Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 assessment tool.

Table 2. Baseline characteristics of the included studies (n=10)

Author, year	Country	Sample size (n)	Age, mean±SD (years)	Sex (n)		Cut-off values (cm)	TP (n)	FP (n)	FN (n)	TN (n)	AP (n)	AN (n)	Sensitivity	Specificity
				Male	Female									
Meko *et al.*, 1995 [28]	United States of America	90	50±N/A	79	11	≥3	12.21	0.00	6.79	71.00	19.00	71.00	0.64	1
Pinchot *et al.*, 2009 [31]	United States of America	155	53.00±1.30	47	108	≥4	14.07	0.00	6.93	132.00	21.00	132.00	0.67	1
Yoon *et al.*, 2011 [25]	South Korea	661	N/A	117	544	≥3	71.59	82.77	2.44	504.23	74.00	587.00	0.97	0.86
Kim *et al.*, 2014 [15]	South Korea	263	45.60±15.50	80	183	≥4	84.98	2.20	7.02	64.80	92.00	67.00	0.92	0.97
Ucler *et al.*, 2015 [30]	Turkey	267	44.40±11.90	N/A	N/A	≥3	4.33	16.56	1.67	73.44	6.00	90.00	0.72	0.82
Kulstad *et al.*, 2016 [32]	South Korea	198	50.80±17.60	57	141	≥4	12.80	24.60	3.20	98.40	16.00	123.00	0.80	0.80
Raguin *et al.*, 2017 [26]	France	843	N/A	207	636	>3	48.15	22.71	37.83	734.30	85.99	757.01	0.56	0.97
Bozbıyık *et al.*, 2017 [29]	Turkey	127	47.80±N/A	38	89	>4	1.66	11.10	1.33	62.90	3.00	74.00	0.55	0.85
Karadeniz *et al.*, 2019 [13]	Turkey	65	N/A	N/A	N/A	>4	3.90	0.00	22.10	39.00	26.00	39.00	0.15	1
Zufry *et al.*, 2023 [27]	Indonesia	83	N/A	69	14	>4	3.85	10.03	5.14	50.97	9.00	61.00	0.43	0.83

AN: actual negative; AP: actual positive; FN: false negative; FP: false positive; N/A: not available; TN: true negative; TP: true positive

Goodness-of-fit and bivariate distribution

Based on the goodness-of-fit analysis, the data exhibited symmetry around the mean, with the left and right sides of the distribution being approximately mirror images, suggesting that the data is well-behaved and supports the assumption of normality (**Figure 3A**). Additionally, bivariate normality test further reinforces this finding by indicating that the data adheres to a normal distribution (**Figure 3B**). These results collectively confirmed the suitability of the meta-analysis model for the data, ensuring that the underlying assumptions for accurate estimation are met.

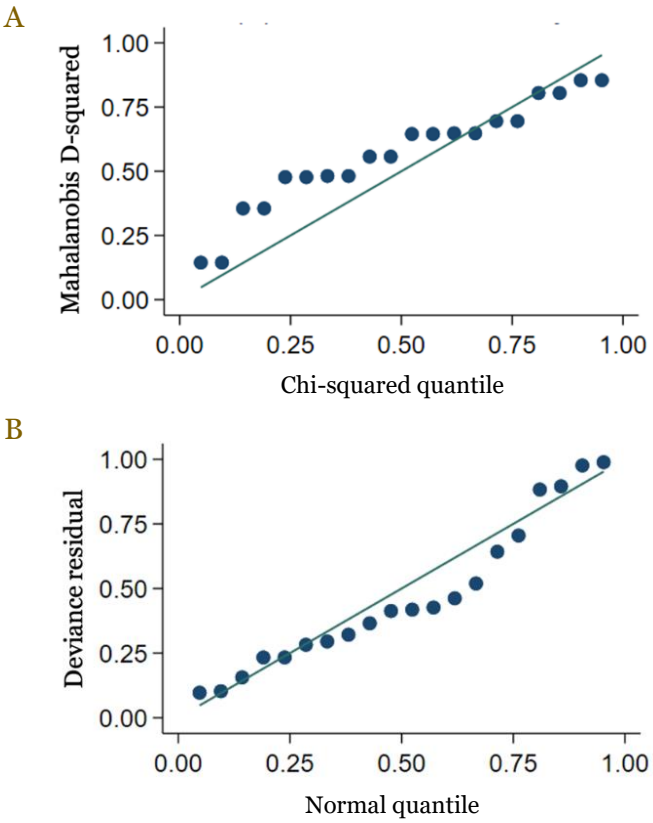

Figure 3. Plots for goodness-of-fit (A) and bivariate normality test (B). (A) Goodness-of-fit analysis: Mahalanobis D-squared statistic and Chi-squared quantile test assess the data's symmetry and distribution. (B) Bivariate normality test: The plot illustrates the relationship between variables, supporting the normal distribution assumption and confirming the model's accuracy.

Diagnostic performance of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules

Sensitivity and specificity of the overall diagnostic performance of US-FNAB were 72% (95%CI: 50–86%; p=0.00) and 96% (95%CI: 87–90%; p=0.00), respectively (**Figure 4**). High heterogeneity was found in both sensitivity (I^2=90.77%; p-Het<0.01) and specificity (I^2=92.48%; p-Het<0.01) pooled estimates (**Figure 4**). The positive likelihood ratio was 16.1 (95%CI: 5.50–47.30), while its negative likelihood ratio was 0.30 (95%CI: 0.16–0.57) (**Table 3**).

Table 3. Summary estimates of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) diagnostic values in large thyroid nodules

Parameters	Estimate	95% confidence interval
Sensitivity	0.72	0.50–0.86
Specificity	0.96	0.87–0.99
Positive likelihood ratio	16.10	5.50–47.30
Negative likelihood ratio	0.30	0.16–0.57
Diagnostic odds ratio	54.00	15.00–188.00

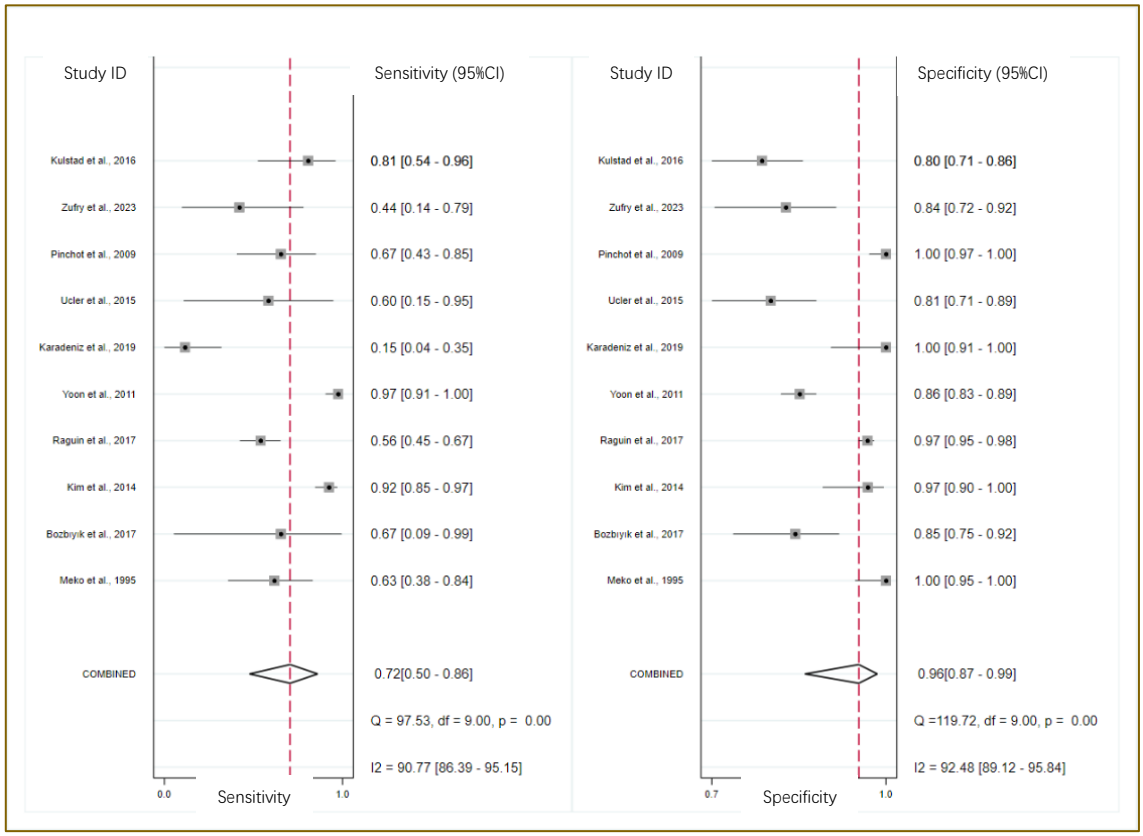

Figure 4. Forest plots for sensitivity and specificity of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules.

Summarized receiver operating characteristic (sROC) of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules

The sROC showed an sAUC of 93%, indicating high diagnostic accuracy (sAUC>75%) (**Figure 5**). The predictive value was calculated based on probability modifying plots (**Figure 6**). The PPV was higher than the NPV, suggested by the shape of the plot, which has more area between the positive test result and the regression line than that between the negative test result and the regression line. The PPV and NPV were found to be 93% (95% CI: 89–98%) and 75% (95% CI: 72–79%), respectively (**Figure 6**).

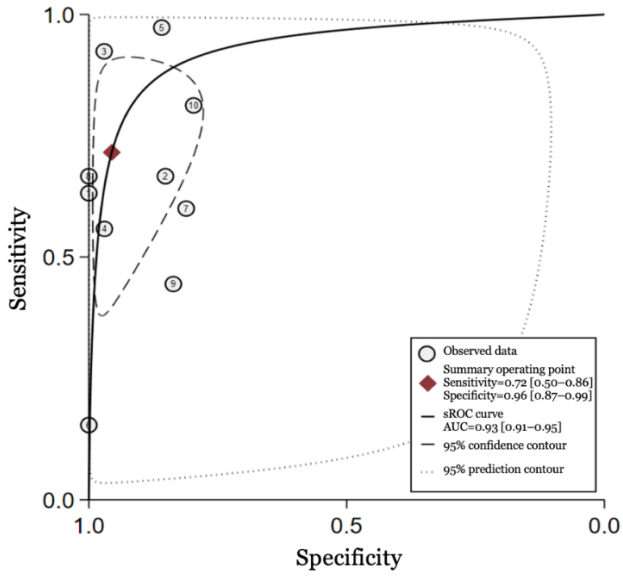

Figure 5. Summarized receiver operating characteristic (sROC) curve demonstrates the diagnostic accuracy of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules.

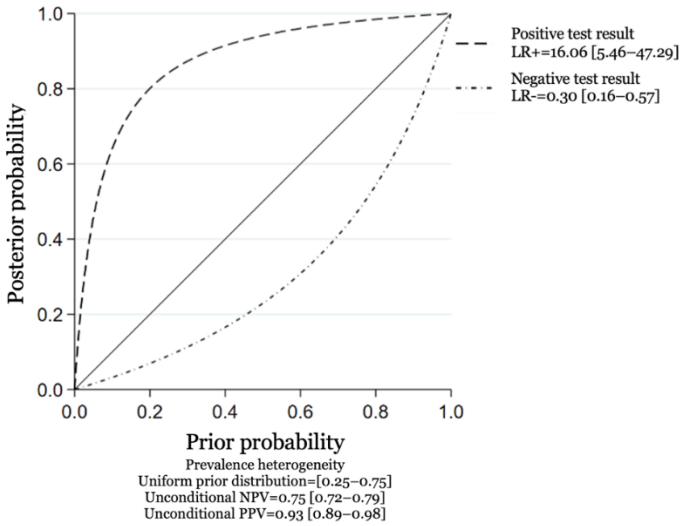

Figure 6. Probability modifying plot illustrates the predictive values (positive predictive value (PPV) and negative predictive value (NPV)) of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules.

Influence of size-effect and publication bias

Cook distance and outlier detection plots were performed, indicating that no study had a disproportionate impact on the overall estimate (**Figure 7A**), which was further confirmed by the outlier detection analysis (**Figure 7B**). Deeks' funnel plot was used to assess publication bias, the plot was found to be statistically symmetrical (p=0.27), suggesting the absence of significant publication bias (**Figure 7C**). Collectively, these findings support the reliability of the pooled estimate.

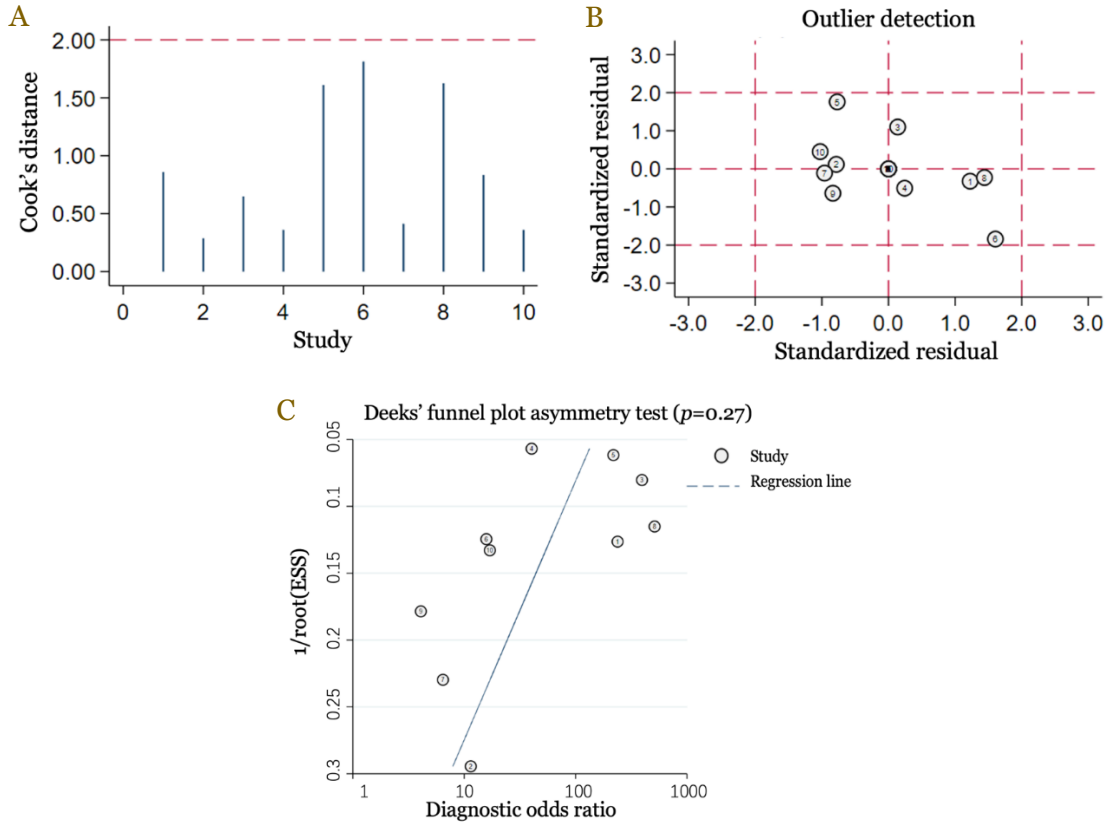

Figure 7. Cook distance (A) and outlier detection plots (B) of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules. Deeks' funnel plot of diagnostic odds ratio of US-FNAB in distinguishing malignancy in large thyroid nodules (C).

Meta-regression

Meta-regression was performed to assess the effect of variables (number of samples, age, male-to-female ratio, thyroid size cut-off, and country) on the diagnostic performance of US-FNAB. The results suggest that factors such as the number of samples, country (high-income and upper-middle income), demographic characteristics (age and sex), and different cut-off values do not significantly influence the diagnostic performance (sensitivity and specificity) of US-FNAB in distinguishing malignancy in large thyroid nodules (**Table 4**).

Table 4. Results from meta-regression assessing the effect number of samples, age, male-to-female ratio, thyroid size cut-off, and country on the diagnostic performance of ultrasonography-guided fine-needle aspiration biopsy (US-FNAB) in distinguishing malignancy in large thyroid nodules.

Variables	Studies (n)	Sensitivity (95%CI)	*p*-sens	Specificity (95%CI)	*p*-spec
Number of samples	10	0.71 (0.50–0.85)	1.00	0.96 (0.87–0.99)	1.00
Age	10	0.79 (0.70–0.86)	0.54	0.98 (0.84–1.00)	0.78
Male-to-female ratio	8	0.74 (0.55–0.87)	0.71	0.96 (0.86–0.99)	0.84
Male>50%	2	0.56 (0.18–0.94)	0.12	0.96 (0.88–1.00)	0.20
Male<50%	6	0.84 (0.71–0.96)		0.95 (0.88–1.00)	
Thyroid size cut-off					
≥4 cm	6	0.67 (0.40–0.93)	0.51	0.96 (0.90–1.00)	0.52
≥3 cm	4	0.78 (0.54–1.00)		0.95 (0.87–1.00)	
Country					
High-income	6	0.82 (0.68–0.96)	0.05	0.96 (0.92–1.00)	0.33
Upper-middle income	4	0.41 (0.10–0.72)		0.91 (0.79–1.00)	

p-sens: *p*-value for sensitivity; *p*-spec: *p*-value for specificity

Discussion

The present study confirms the reliability of US-FNAB in distinguishing malignancy in large thyroid nodules, with an overall diagnostic sensitivity of 72% (95%CI: 50–86%; p=0.00) and specificity of 96% (95%CI: 87–90%; p=0.00). The PPV and NPV were 93% (95%CI: 89–98%) and 75% (95%CI: 72–79%), respectively, with a sAUC of 93%, indicating high diagnostic accuracy. Meta-regression analysis revealed that factors such as the number of samples, country (high-income vs upper-middle income), demographic characteristics (age and sex), and different thyroid size cut-off values did not significantly impact the sensitivity or specificity of US-FNAB.

A prevailing concern with large thyroid nodules is the higher risk of malignancy, leading to a pervasive stigma associated with large thyroid nodules diagnosis. Subsequently, many studies have advocated for thyroidectomy as a precautionary measure for large thyroid nodules, irrespective of US-FNAB results [10,15,18,31,33-37]. However, the present study challenges this recommendation by underscoring the effectiveness of US-FNAB in accurately identifying malignancy in large thyroid nodules, thus minimizing the need for unnecessary thyroidectomy. Furthermore, US-FNAB accuracy varies with thyroid nodules, and a study showed that the accuracy of US-FNAB was higher in large thyroid nodules with US features suspicious of malignancy, such as a solid component, ill-defined margin, hypoechogenicity or marked hypoechogenicity, or any calcifications (micro- or macro-) [15]. By identifying more susceptible patients through US-FNAB screening, the present study proposes a nuanced approach to managing large thyroid nodules, enabling clinicians to minimize invasive procedures in low-risk patients while identifying those at a higher risk of malignancy [38].

Bethesda System for Reporting Thyroid Cytopathology is widely accepted in cytopathology practice worldwide and has established a standardized six-tiered system for the stratification of thyroid nodules [22,39]. The 2023 European Thyroid Association guideline recommends thyroidectomy for symptomatic nodules, those initially classified as benign but later becoming symptomatic, elevated calcitonin levels, responsive calcitonin in *RET*-mutated carriers, and indeterminate or malignant cytology (Bethesda class III, IV, V, VI) [40]. However, there is no specified size cutoff for large thyroid nodules eligible for thyroidectomy. On the other hand, the 2015 American Thyroid Association guideline recommends total thyroidectomy for indeterminate nodules with increased malignancy risk, including those cytologically suspicious,

positive for specific mutations, sonographically suspicious, or large (>4 cm) [41]. To the best of our knowledge, currently, there is no specific guideline for managing large thyroid nodules. Subsequently, its development in the near future is needed.

In the present study, the overall pooled estimate exhibited high heterogeneity (I^2=98%, p-Het<0.01). This variability may stem from poor inter- and intra-observer agreement associated with US-FNAB among practitioners. However, studies included in the present meta-analysis did not report any information on intraobserver variability, a key gap that could affect diagnostic consistency and accuracy. Ultrasound examination's subjectivity adds another layer of complexity, introducing inconsistencies between different examiners (inter-observer variability) and even for the same examiner (intra-observer variability) [42]. A previous study found significant intra- and inter-observer variability among pathologists and cytologists in thyroid US-FNAB analysis [36]. Pathologists showed moderate-to-substantial intra-observer agreement in evaluating non-benign thyroid biopsies, but inter-observer agreement was below the acceptable limit when using cytopathologists as a reference [43]. Moreover, the influence of healthcare center volume on the inter-observer agreement is also notable, as high-volume and low-volume centers can impact results [37]. A study revealed more frequent changes in US-FNAB results for nodules at intermediate/high risk compared to no/low-risk nodules [44].

The interpretation of US-FNAB poses challenges due to the limited morphological differences between non-neoplastic and neoplastic thyroid conditions and the variability in US-FNAB specimen preparation and interpretation [45]. A previous study found substantial inter- and intra-observer variability in the cytopathologic and histopathologic evaluation of thyroid nodules, with inter-observer disagreement rates ranging from 11% to 35% [46]. Thyroid US-FNAB is traditionally performed by various practitioners, including endocrinologists, surgeons, radiologists, and cytopathologists, leading to varying specimen quality [45]. Cytologic interpretation variations can arise from individual observers having specific diagnostic biases, which may be due to not strictly applying requested criteria or receiving criteria lacking adequate detail or descriptiveness to encompass all potential diagnostic possibilities [45].

The most challenging category in the Thyroid Bethesda System is class III/atypia of unknown significance/follicular lesion of undetermined significance [45,47]. For instance, detecting the follicular variant of papillary thyroid carcinoma (PTC) leads to some thyroid aspirates being classified as 'atypia of undetermined significance/follicular lesion of undetermined significance' (AUS/FLUS) [48]. A study suggested that macrocalcification in thyroid nodules indicates a high risk of PTC, and combining US-FNAB with *BRAF V600E* enhances the identification of macro-calcified thyroid nodules [49].

Histopathological changes in thyroid nodules may evolve over time, potentially affecting the diagnostic accuracy of initial assessments [50]. Therefore, serial evaluations are recommended to confirm US-FNAB results and monitor for any progressive changes in the lesion. In cases requiring further confirmation, core-needle biopsy (CNB) can serve as a complementary diagnostic method. A meta-analysis reported FNAB with 72% sensitivity and 99% specificity, while CNB showed higher sensitivity (83%) and similar specificity (99%). The sAUC was 0.9025 for FNAB and 0.7926 for CNB, with no significant difference between them (p=0.164) [51]. These findings highlighted the value of integrating both diagnostic techniques. While US-FNAB remains a reliable first-line approach, CNB offers enhanced sensitivity and serves as a complementary method for confirming malignancy, particularly in nodules with indeterminate or inconclusive FNAB results [51].

To the best of our knowledge, the present study is the first review investigating the diagnostic accuracy of preoperative US-FNAB in comparison to posthistopathological tests. However, it is essential to acknowledge certain limitations within the present study. High heterogeneity rates among the included studies were observed, probably attributed to poor inter- and intra-observer agreement among pathologists. Despite this, the present study attempted to mitigate systematic errors by conducting sensitivity analyses, which did not reveal significant differences from the original analyses. Furthermore, the goodness-of-fit analysis confirmed that the data met the assumptions necessary for accurate meta-analysis estimation. Although these measures mitigate some concerns, the observed heterogeneity remains a significant limitation.

The present study has successfully synthesized the diagnostic accuracy of FNAB in detecting malignancy among large thyroid nodules using data from different countries with high quality. Furthermore, no publication bias was not detected. However, the present systematic review was restricted to a limited number of databases, which may have impacted the comprehensiveness of the literature search and potentially excluded relevant studies not indexed in the accessible databases. Furthermore, the present study did not contact experts in this field to ask for possible data that have not been published. The absence of information on intraobserver variability in the studies included in the present meta-analysis represents a notable limitation, as variability within a single observer could influence the consistency and reliability of the results over time. Additionally, the lack of data on interobserver agreement between pathologists and ultrasound operators is significant, as differences in interpretation between observers may contribute to inconsistencies in diagnostic outcomes.

In interpreting the results, readers were encouraged to pay attention on the high heterogeneity stemming from the poor inter- and intra-observer agreement in FNAB readings. Additional research is warranted, including implementation of prospective studies with a well-defined cohort of patients to establish an up-to-date and comprehensive dataset for analysis. Furthermore, conducting long-term follow-up studies to evaluate the accuracy of US-FNAB in predicting malignancy over an extended period might enhance the understanding of malignancy in large thyroid nodules, considering the potential for changes in nodule characteristics.

Conclusion

US-FNAB demonstrated high diagnostic accuracy in distinguishing malignancy in large thyroid nodules, with a sensitivity of 72%, specificity of 96%, PPV of 93%, and NPV of 75%, supported by an sAUC of 93%, emphasizing its role in reducing unnecessary thyroidectomy by identifying high-risk patients, and challenging the conventional practice of routine thyroidectomy for large thyroid nodules. Sample size, demographic factors, country income level, and thyroid size cut-offs have no significant effect on US-FNAB diagnostic performance. Serial evaluations and complementary diagnostic methods are recommended to confirm malignancy in large thyroid nodules, particularly in nodules with indeterminate or inconclusive US-FNAB results.

Ethics approval

Not required.

Acknowledgments

Not applicable.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Declaration of artificial intelligence use

This study used artificial intelligence (AI) tools and methodologies of which AI-based language model, ChatGPT, was employed in the language refinement (improving grammar, sentence structure, and readability of the manuscript). We confirm that all AI-assisted processes were critically reviewed by the authors to ensure the integrity and reliability of the results. The final decisions and interpretations presented in this article were solely made by the authors.

How to cite

Zulfa PO, Iqhramullah M, Zufry H. Diagnostic accuracy of preoperative ultrasonography-guided fine-needle aspiration biopsy in distinguishing malignancy in large thyroid nodules: A systematic review, meta-analysis, and meta-regression. Narra J 2025; 5 (1): e1120 - http://doi.org/10.52225/narra.v5i1.1120.

References

1. Turkkan E, Uzum Y. Evaluation of thyroid nodules in patients with fine-needle aspiration biopsy. Cureus 2023;15(9):e44569.
2. Brito JP, Yarur AJ, Prokop LJ, *et al.* Prevalence of thyroid cancer in multinodular goiter versus single nodule: A systematic review and meta-analysis. Thyroid 2013;23(4):449-455.
3. Mu C, Ming X, Tian Y, *et al.* Mapping global epidemiology of thyroid nodules among general population: A systematic review and meta-analysis. Front Oncol 2022;12:1029926.
4. Giovanella L, Campennì A, Tuncel M, *et al.* Integrated diagnostics of thyroid nodules. Cancers (Basel) 2024;16(2):311.
5. Cantisani V, De Silvestri A, Scotti V, *et al.* US-Elastography with different techniques for thyroid nodule characterization: Systematic review and meta-analysis. Front Oncol 2022;12:845549.
6. Bernet VJ, Chindris AM. Update on the evaluation of thyroid nodules. J Nucl Med 2021;62 Suppl 2:13S-19S.
7. Li M, Dal Maso L, Vaccarella S. Global trends in thyroid cancer incidence and the impact of overdiagnosis. Lancet Diabetes Endocrinol 2020;8(6):468-470.
8. Ugurluoglu C, Dobur F, Karabagli P, *et al.* Fine needle aspiration biopsy of thyroid nodules: Cytologic and histopathologic correlation of 1096 patients. Int J Clin Exp Pathol 2015;8(11):14800-14805.
9. Kim HK, Kim SY, Lee YS, *et al.* Suspicious thyroid nodules 4 cm require a diagnostic lobectomy regardless of their benign fine needle aspiration results. Asian J Surg 2022;45(5):1113-1116.
10. Kang S, Kim E, Lee S, *et al.* Do large thyroid nodules (⩾4 cm) without suspicious cytology need surgery? Front Endocrinol (Lausanne) 2023;14:1252503.
11. Pantanowitz L, Thompson LDR, Jing X, *et al.* Is thyroid core needle biopsy a valid compliment to fine-needle aspiration? J Am Soc Cytopathol 2020;9(5):383-388.
12. Hahn SY, Shin JH, Oh YL, *et al.* Comparison between fine needle aspiration and core needle biopsy for the diagnosis of thyroid nodules: Effective indications according to US findings. Sci Rep 2020;10(1):4969.
13. Karadeniz E, Yur M, Temiz A, *et al.* Malignancy risk for thyroid nodules larger than 4 cm and diagnostic reliability of ultrasound-guided FNAB results. Turk J Surg 2019;35(1):13–18.
14. Magister MJ, Chaikhoutdinov I, Schaefer E, *et al.* Association of thyroid nodule size and Bethesda class with rate of malignant disease. JAMA Otolaryngol Head Neck Surg 2015;141(12):1089-1095.
15. Kim JH, Kim NK, Oh YL, *et al.* The validity of ultrasonography-guided fine needle aspiration biopsy in thyroid nodules 4 cm or larger depends on ultrasonography characteristics. Endocrinol Metab (Seoul) 2014;29(4):545-552.
16. Cavallo A, Johnson DN, White MG, *et al.* Thyroid nodule size at ultrasound as a predictor of malignancy and final pathologic size. Thyroid 2017;27(5):641-650.
17. Al-Hakami HA, Alqahtani R, Alahmadi A, *et al.* Thyroid nodule size and prediction of cancer: A study at tertiary care hospital in Saudi Arabia. Cureus 2020;12(3):e7478.
18. McCoy KL, Jabbour N, Ogilvie JB, *et al.* The incidence of cancer and rate of false-negative cytology in thyroid nodules greater than or equal to 4 cm in size. Surgery 2007;142(6):837-844.e3.
19. Cao M, Yu T, Miao X, *et al.* The preferred surgical choice for intermediate-risk papillary thyroid cancer: Total thyroidectomy or lobectomy? A systematic review and meta-analysis. Int J Surg 2024;110(8):5087-5100.
20. Page MJ, McKenzie JE, Bossuyt PM, *et al.* The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
21. van de Schoot R, de Bruin J, Schram R, *et al.* An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021;3(2):125-133.
22. Cibas ES, Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. Thyroid 2017;27(11):1341-1346.
23. Whiting PF, Rutjes AW, Westwood ME, *et al.* QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-536.
24. Dwamena BA, Sylvester R, Carlos RC. MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies. RePEc 2009;2:2-25.

25. Yoon JH, Kwak JY, Moon HJ, *et al.* The diagnostic accuracy of ultrasound-guided fine-needle aspiration biopsy and the sonographic differences between benign and malignant thyroid nodules 3 cm or larger. Thyroid 2011;21(9):993-1000.
26. Raguin T, Schneegans O, Rodier J, *et al.* Value of fine-needle aspiration in evaluating large thyroid nodules. Head Neck 2017;39(1):32-36.
27. Zufry H, Nazaruddin N, Zulfa PO, *et al.* Comparative analysis of accuracy between fine-needle aspiration biopsy and postoperative histopathology for detecting large thyroid nodules: A retrospective observational study. Narra J 2023;3(2):e206.
28. Meko JB, Norton JA. Large cystic/solid thyroid nodules: A potential false-negative fine-needle aspiration. Surgery 1995;118(6):996-1004.
29. Bozbıyık O, Öztürk Ş, Ünver M, *et al.* Reliability of fine needle aspiration biopsy in large thyroid nodules. Turk J Surg 2017;33(1):10-13.
30. Ucler R, Usluogulları CA, Tam AA, *et al.* The diagnostic accuracy of ultrasound-guided fine-needle aspiration biopsy for thyroid nodules three centimeters or larger in size. Diagn Cytopathol 2015;43(8):622-628.
31. Pinchot SN, Al-Wagih H, Schaefer S, *et al.* Accuracy of fine-needle aspiration biopsy for predicting neoplasm or carcinoma in thyroid nodules 4 cm or larger. Arch Surg 2009;144(7):649-655.
32. Kulstad R. Do all thyroid nodules >4 cm need to be removed? An evaluation of thyroid fine-needle aspiration biopsy in large thyroid nodules. Endocr Pract 2016;22(7):791-798.
33. Wharry LI, McCoy KL, Stang MT, *et al.* Thyroid nodules (⩾4 cm): Can ultrasound and cytology reliably exclude cancer? World J Surg 2014;38(3):614-621.
34. Carrillo JF, Frias-Mendivil M, Ochoa-Carrillo FJ, *et al.* Accuracy of fine-needle aspiration biopsy of the thyroid combined with an evaluation of clinical and radiologic factors. Otolaryngol Head Neck Surg 2000;122(6):917-921.
35. Giles WH, Maclellan RA, Gawande AA, *et al.* False negative cytology in large thyroid nodules. Ann Surg Oncol 2015;22(1):152-157.
36. Koo DH, Song K, Kwon H, *et al.* Does tumor size influence the diagnostic accuracy of ultrasound-guided fine-needle aspiration cytology for thyroid nodules? Int J Endocrinol 2016;2016:3803647.
37. Bestepe N, Ozdemir D, Tam AA, *et al.* Malignancy risk and false-negative rate of fine needle aspiration cytology in thyroid nodules ⩾4.0 cm. Surgery 2016;160(2):405-412.
38. Vargas-Uricoechea H, Meza-Cabrera I, Herrera-Chaparro J. Concordance between the TIRADS ultrasound criteria and the BETHESDA cytology criteria on the nontoxic thyroid nodule. Thyroid Res 2017;10:1.
39. Guerreiro SC, Tastekin E, Mourao M, *et al.* Impact of the 3rd edition of the Bethesda system for reporting thyroid cytopathology on grey zone categories. Acta Cytol 2023;67(6):593-603.
40. Durante C, Hegedüs L, Czarniecka A, *et al.* 2023 European Thyroid Association clinical practice guidelines for thyroid nodule management. Eur Thyroid J 2023;12(5):e230067.
41. Haugen BR, Alexander EK, Bible KC, *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):1-133.
42. Dobruch-Sobczak K, Adamczewski Z, Dedecjus M, *et al.* Summary of meta-analyses of studies considering lesion size cut-off thresholds for the assessment of eligibility for FNAB and sonoelastography and inter- and intra-observer agreement in estimating the malignant potential of focal lesions of the thyroid gland. J Ultrason 2022;22(89):130-135.
43. Kuzan TY, Güzelbey B, Turan Güzel N, *et al.* Analysis of intra-observer and inter-observer variability of pathologists for non-benign thyroid fine needle aspiration cytology according to Bethesda system categories. Diagn Cytopathol 2021;49(7):850-855.
44. Scappaticcio L, Trimboli P, Iorio S, *et al.* Repeat thyroid FNAC: Inter-observer agreement among high- and low-volume centers in Naples metropolitan area and correlation with the EU-TIRADS. Front Endocrinol (Lausanne) 2022;13:1001728.
45. Kocjan G, Chandra A, Cross PA, *et al.* The interobserver reproducibility of thyroid fine-needle aspiration using the UK Royal College of Pathologists' classification system. Am J Clin Pathol 2011;135(6):852-859.
46. Cibas ES, Baloch ZW, Fellegara G, *et al.* A prospective assessment defining the limitations of thyroid nodule pathologic evaluation. Ann Intern Med 2013;159(5):325-332.
47. Kurian EM, Dawlett M, Wang J, *et al.* The triage efficacy of fine needle aspiration biopsy for follicular variant of papillary thyroid carcinoma using the Bethesda reporting guidelines. Diagn Cytopathol 2012;40 Suppl 1:E69-E73.

48. Ono JC, Wilbur DC, Lee H, *et al.* Cytologic features of focal papillary thyroid carcinoma arising within follicular adenoma: A masked cytomorphologic analysis of 17 cases. Acta Cytol 2011;55(6):531-538.
49. Ye M, Wu S, Zhou Q, *et al.* Association between macrocalcification and papillary thyroid carcinoma and corresponding valuable diagnostic tool: Retrospective study. World J Surg Oncol 2023;21(1):149.
50. Fatima T, Ali S, Rawtani G, *et al.* Age related cytoarchitectural comparison of histopathological changes in thyroid nodule. J Bahria Univ Med Dent Coll 2023;13(02):135-139.
51. Lan L, Luo Y, Zhou M, *et al.* Comparison of diagnostic accuracy of thyroid cancer with ultrasound-guided fine-needle aspiration and core-needle biopsy: A systematic review and meta-analysis. Front Endocrinol (Lausanne) 2020;11:44.