Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Effect of Artificial Intelligence Assistance on Diagnostic Performance and Interobserver Agreement among Intermediate-Experience Physicians in Thyroid Nodule Ultrasonography

Bekjon Sattarov
Tashkent State Medical University
Hee Jung Suh
Severance Hospital, Yonsei University College of Medicine,
In Hye Chae
Severance Hospital, Yonsei University College of Medicine,
Sun Huh
Yonsei National Health Internal Medicine Clinic,
Sung Hoon Hong
Soonchunhyang University Hospital Cheonan
Kim Ho Jung
Kim Ho Jung ENT Clinic
Hayoung Kim
Severance Hospital, Yonsei University College of Medicine,
Jin Young Kwak
docjin@yuhs.ac

Severance Hospital, Yonsei University College of Medicine,

Research Article

Keywords: Ultrasonography, Thyroid nodule, Artificial intelligence, Computer-assisted diagnosis, Interobserver agreement

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11144213/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose

To evaluate the effect of artificial intelligence (AI) assistance on the diagnostic performance and interobserver agreement of physicians with intermediate-experience in the ultrasonographic (US) assessment of thyroid nodules.

Methods

This retrospective study included 423 thyroid nodules (150 benign and 273 malignant) with cytological or histopathological diagnoses. Five physicians with intermediate experience independently classified nodules as benign or malignant using representative US images, first without and subsequently with AI assistance. Diagnostic performance metrics—including sensitivity, specificity, and accuracy—were calculated with corresponding 95% confidence intervals. Interobserver agreement was assessed using Fleiss' kappa, and reader perceptions of AI usefulness were recorded using a three-level categorical scale.

Results

AI assistance was associated with modest but consistent directional improvements in diagnostic accuracy across all readers, although none reached statistical significance. Sensitivity increased in four readers and specificity in two, but these differences were not statistically significant. Interobserver agreement improved from $\kappa = 0.653$ without AI assistance to $\kappa = 0.709$ with AI assistance. Most readers perceived AI as "somewhat helpful" despite limited quantitative performance gains.

Conclusion

AI assistance did not significantly improve the diagnostic performance of physicians with intermediate experience in the US assessment of thyroid nodules. Interobserver agreement was numerically higher with AI, but the significance of this difference requires further evaluation.

Key points

AI assistance did not significantly improve the diagnostic performance of physicians with intermediate experience in the US assessment of thyroid nodules.

Interobserver agreement improved from $\kappa = 0.653$ to $\kappa = 0.709$ with AI assistance.

AI may support diagnostic standardization among physicians with intermediate US experience.

Introduction

Thyroid nodules are a common clinical finding, detected by high-resolution ultrasonography (US) in approximately 19%–68% of the adult population (1–3). Although most thyroid nodules are benign, a small proportion exhibit malignant characteristics, requiring accurate evaluation and effective clinical management (2–4). Ultrasonography remains the primary imaging modality for initial evaluation due to its accessibility, non-invasiveness, and cost-effectiveness (3–7). To standardize thyroid US reporting and improve diagnostic consistency, several classification systems have been introduced (5,6,8–11). However, US interpretation remains subjective and highly operator-dependent, with variability among observers (12,13). Consequently, even experienced radiologists achieve only moderate interobserver agreement (12,14,15).

In recent years, artificial intelligence (AI), particularly deep-learning (DL)-based systems, has emerged as a valuable adjunct in thyroid imaging (7,16–18). These systems automatically analyze US images, estimate malignancy risk, and assist in biopsy decision-making (7,17,18). Commercially available AI platforms such as Koios DS, S-Detect, and AmCAD-UT have demonstrated diagnostic performance comparable to expert radiologists, offering substantial clinical utility (16,18,19). Recent multicenter studies and clinical reports indicate that AI-assisted systems significantly enhance diagnostic sensitivity and accuracy, particularly among less experienced clinicians (19,20), whereas no statistically significant improvement has been consistently observed in highly experienced physicians (18). Nevertheless, caution is required when integrating AI into routine clinical practice. Diagnostic outcomes can vary depending on clinician expertise, quality of US acquisition, differing patient populations, and healthcare environments (7,16,19,21). Studies suggest greater performance gains for clinicians with less experience, highlighting variability in AI's impact based on clinician skill and clinical context (18–20).

To date, many AI studies in thyroid imaging have been conducted in controlled research environments, often using retrospective static-image datasets and including results from both expert and less experienced physicians (13,20,22,23). Therefore, this study aims to fill this critical knowledge gap by evaluating diagnostic performances for thyroid nodule US assessment among intermediate-experience clinicians with and without AI assistance. Additionally, we explore the practicality of implementing AI as a clinical decision support tool in environments where subspecialized radiological expertise is limited.

Materials and methods

Study Population

In this retrospective study, 975 consecutive patients who underwent fine-needle aspiration (FNA) or staging ultrasonography (US) for thyroid nodules at a tertiary referral center between February and April 2019 were initially screened. Of the initial cohort, 588 thyroid nodules measuring ≥ 1 cm were identified. Nodules were excluded if they lacked definitive cytological or histopathological diagnosis, had nondiagnostic or indeterminate cytology results, or did not receive subsequent management (such as

repeat FNA or surgery) following an initial nondiagnostic FNA result. In addition, nodules aspirated by operators with less than one year of dedicated thyroid imaging experience were also excluded. Ultimately, 423 nodules with conclusive diagnoses confirmed by either FNA cytology or surgical pathology were included in the final analysis (Fig. 1). All US examinations were performed using a 5–18 MHz linear transducer (EPIQ Elite Ultrasound System, Philips Healthcare, Andover, MA, USA) by one of 5 radiologists dedicated to thyroid imaging, each with 2–20 years of experience.

Ultrasound Acquisition and Assessment

A board-certified radiologist with 20 years of experience in thyroid imaging, who was blinded to clinical and pathologic information, selected a representative US image for each thyroid nodule from the PACS and exported it as a JPEG file. The same radiologist then manually delineated a square ROI covering the entire target nodule for each image using Microsoft Paint (Windows 10). In this study, five readers performed a dichotomous assessment of thyroid nodules as malignant or benign. The clinical experience of these readers was specifically defined as having performed or interpreted between 300 and 1,000 thyroid US cases (24). The readers consisted of three board-certified radiologists, one endocrine surgeon, and one otorhinolaryngologist. Although TI-RADS–based analyses were performed in a subset of cases, TI-RADS classification was not applied consistently by all readers across the entire dataset; therefore, this work should not be described as a TI-RADS–based analysis. All readers first evaluated the nodules without AI assistance and then reassessed the same cases with AI support, providing binary judgments of malignancy.

For this study, an in-house AI-based computer-aided diagnosis program was used (23). The program provided malignancy risk scores ranging from 0 to 100% and categorization according to the TI-RADS for each nodule. The AI model was trained and validated on a large, institutionally curated dataset and has previously shown performance comparable to experienced radiologists (23).

To evaluate the perceived utility of AI assistance, the five readers independently rated how helpful the AI output was for their diagnosis using a three-level categorical scale (very helpful, somewhat helpful, and not helpful).

Statistical Analysis

For each thyroid nodule, the reference standard was defined as either a definitive histopathological diagnosis or cytological diagnosis, as documented in the medical records. Continuous variables were analyzed using the Mann–Whitney U test, and categorical variables were analyzed using the chi-square test or Fisher’s exact test, as appropriate. Normality of continuous variables was assessed using the Kolmogorov–Smirnov test.

Sensitivity, specificity, and accuracy were calculated for interpretations performed without and with AI assistance and were reported with corresponding 95% confidence intervals (CIs). Comparisons between

interpretations without and with AI assistance were performed using the McNemar test for paired proportions. The distribution of reader-rated perceived usefulness of AI assistance was analyzed using Pearson's chi-square test. Interobserver agreement among the five readers was assessed using Fleiss' kappa coefficient, with 95% confidence intervals calculated separately for interpretations performed without and with AI assistance.

All statistical analyses were performed using SPSS software (version 25.0; IBM Corp., Armonk, NY, USA). A p-value less than 0.05 was considered statistically significant.

Results

Patient demographics are summarized in Table 1. A total of 423 nodules from 423 patients were included in the analysis, comprising 150 benign and 273 malignant nodules. The median age was 53 years (range, 23–87) in the benign group and 41 years (range, 19–81) in the malignant group ($p < 0.001$). The median tumor size was 22 mm (range, 10–65) vs. 14 mm (range, 10–54), respectively ($p < 0.001$). No statistically significant difference in gender distribution was observed ($p = 0.456$).

Table 1
Patient demographics

Variable	Total (n = 423)	Benign (n = 150)	Malignant (n = 273)	p value
Age, years	45 (19–87)	53 (23–87)	41 (19–81)	< 0.001
Gender				0.456
Male	99 (23.40)	32 (21.33)	67 (24.54)	
Female	324 (76.60)	118 (78.67)	206 (75.46)	
Tumor size, mm	16 (10–65)	22 (10–65)	14 (10–54)	< 0.001

Data are presented as n (%) or median (range).

The diagnostic performance of the five readers with and without AI assistance is summarized in Table 2. All five readers showed modest gains in accuracy after receiving AI assistance, although this improvement was not statistically significant (Reader 1: 83.92% [95% CI: 80.42–87.42] without AI vs 84.40% [80.94–87.86] with AI; Reader 2: 81.09% [77.36–84.82] vs 81.56% [77.86–85.26]; Reader 3: 78.72% [74.82–82.62] vs 78.96% [75.08–82.84]; Reader 4: 78.96% [75.08–82.84] vs 79.20% [75.33–83.07]; Reader 5: 81.09% [77.36–84.82] vs 82.27% [78.63–85.91]). The impact of AI assistance on sensitivity varied across readers. Sensitivity increased in Readers 2, 4, and 5 (Reader 2: 76.92% [95% CI: 71.92–81.92] without AI vs 78.39% [73.51–83.27] with AI; Reader 4: 76.92% [71.92–81.92] vs 77.29% [72.32–82.26]; Reader 5: 80.22% [75.49–84.95] vs 80.59% [75.90–85.28]). Sensitivity increased slightly in Reader 3 (73.63% [68.40–78.86] vs 73.99% [68.79–79.19]) and slightly decreased in Reader 1 (84.98% [80.74–89.22] vs 84.62% [80.34–88.90]). With respect to specificity, improvements were observed in Readers 1 and 5 (Reader 1: 82.00% [95% CI: 75.85–88.15] without AI vs 84.00% [78.13–89.87] with AI;

Reader 5: 82.67% [76.61–88.73] vs 85.33% [79.67–90.99]). Specificity was unchanged in Readers 3 and 4 (Reader 3: 88.00% [82.80–93.20] in both conditions; Reader 4: 82.67% [76.61–88.73] in both conditions), while it decreased slightly in Reader 2 (88.67% [83.60–93.74] vs 87.33% [82.01–92.65]). The point estimate of interobserver agreement was higher with AI assistance. Fleiss' kappa increased from 0.653 (95% CI, 0.629–0.678) without AI to 0.709 (95% CI, 0.684–0.734) with AI assistance.

Table 2
Diagnostic performance of readers with and without AI assistance

	Method	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Reader 1	without AI assistance	84.98 (80.74 ~ 89.22)	82.00 (75.85 ~ 88.15)	83.92 (80.42 ~ 87.42)
	with AI assistance	84.62 (80.34 ~ 88.90)	84.00 (78.13 ~ 89.87)	84.40 (80.94 ~ 87.86)
	p value*	0.796	0.316	0.683
Reader 2	without AI assistance	76.92 (71.92 ~ 81.92)	88.67 (83.60 ~ 93.74)	81.09 (77.36 ~ 84.82)
	with AI assistance	78.39 (73.51 ~ 83.27)	87.33 (82.01 ~ 92.65)	81.56 (77.86 ~ 85.26)
	p value*	0.247	0.413	0.637
Reader 3	without AI assistance	73.63 (68.40 ~ 78.86)	88.00 (82.80 ~ 93.20)	78.72 (74.82 ~ 82.62)
	with AI assistance	73.99 (68.79 ~ 79.19)	88.00 (82.80 ~ 93.20)	78.96 (75.08 ~ 82.84)
	p value*	0.763	> 0.999	0.782
Reader 4	without AI assistance	76.92 (71.92 ~ 81.92)	82.67 (76.61 ~ 88.73)	78.96 (75.08 ~ 82.84)
	with AI assistance	77.29 (72.32 ~ 82.26)	82.67 (76.61 ~ 88.73)	79.20 (75.33 ~ 83.07)
	p value*	0.842	> 0.999	0.873
Reader 5	without AI assistance	80.22 (75.49 ~ 84.95)	82.67 (76.61 ~ 88.73)	81.09 (77.36 ~ 84.82)
	with AI assistance	80.59 (75.90 ~ 85.28)	85.33 (79.67 ~ 90.99)	82.27 (78.63 ~ 85.91)
	p value*	0.564	0.099	0.095

CI, confidence intervals; AI, artificial intelligence

*p values represent comparisons between results obtained with and without AI assistance.

A representative case in which three readers changed their initial classification from benign to malignant after AI assistance is shown in Fig. 2.

Perceptions of AI assistance differed by reader ($p < 0.001$), with most case-level ratings indicating that AI was “very helpful” or “somewhat helpful.” Reader 2 most frequently classified AI as “very helpful” (87.23%, 369/423), whereas Reader 3 most often categorized AI as “somewhat helpful” (93.85%, 397/423). Reader 4 reported the highest proportion of “not helpful” ratings (13.71%, 58/423) compared with the other readers (Table 3).

Table 3
Reader-rated perceived usefulness of AI assistance

Reader 1	**Very helpful, n (%)**	**Somewhat helpful, n (%)**	**Not helpful at all, n (%)**
	38 (8.98)	**368 (87.00)**	**17 (4.02)**
Reader 2	369 (87.23)	19 (4.49)	35 (8.27)
Reader 3	11 (2.60)	397 (93.85)	15 (3.55)
Reader 4	49 (11.58)	316 (74.70)	58 (13.71)
Reader 5	1 (0.24)	387 (91.49)	35 (8.27)

Discussion

In this study, we evaluated the impact of AI assistance on the diagnostic performance and interobserver agreement of physicians with intermediate experience in the US assessment of thyroid nodules using a dichotomous benign–malignant classification. AI assistance was associated with modest but consistent directional improvements in diagnostic accuracy across all readers; however, these changes did not reach statistical significance. The absence of statistically significant performance gains may reflect the intermediate experience level of the readers. Previous studies have demonstrated that AI-based computer-aided diagnosis systems tend to provide the greatest measurable benefit for less experienced clinicians, whereas improvements are often attenuated among experienced readers (18–20). Physicians with intermediate experience may operate near a performance plateau, limiting the magnitude of observable gains in conventional performance metrics.

Our findings are generally consistent with prior investigations evaluating AI-assisted thyroid US interpretation (7,18,22). Several studies have demonstrated that AI-based CAD can achieve diagnostic performance comparable to that of expert radiologists and may improve performance among less experienced clinicians (18,20,22). Kang et al. reported significant improvements in diagnostic accuracy and sensitivity among junior physicians using CAD assistance, while the benefit was less pronounced among more experienced readers (20). Similarly, Ha et al. showed that AI support improved diagnostic consistency and reduced variability among non-expert clinicians (18). However, relatively few studies have specifically evaluated radiologists with intermediate levels of experience, despite this group

representing a substantial proportion of clinicians in routine practice. By focusing on this experience level, our study highlights the importance of contextualizing AI performance according to the intended user group and clinical environment, rather than assuming uniform benefits across all levels of expertise.

The numerical difference in interobserver agreement is an important descriptive finding, but it should be interpreted cautiously. Ultrasonography remains inherently operator-dependent, and interobserver variability persists even when standardized reporting systems such as TI-RADS are applied (12,14). Although Fleiss' kappa was numerically higher with AI, no paired comparison was performed to establish whether the difference was statistically significant. Further validation is necessary before concluding that AI support improves standardization in routine clinical practice.

An additional noteworthy aspect of our results is the discrepancy between objective diagnostic performance metrics and subjective reader perceptions. Despite minimal changes in accuracy, sensitivity, and specificity, reader perceptions of AI assistance varied across individuals. Overall, AI output was generally rated as either "very helpful" or "somewhat helpful," while a smaller proportion of cases were classified as "not helpful at all," reflecting inter-reader variability in how AI output is incorporated into clinical decision-making. Because diagnostic confidence, interpretation time, and downstream management decisions were not measured, the usefulness ratings should be considered subjective perceptions rather than evidence of workflow or clinical benefit.

Several limitations of this study should be acknowledged. First, the retrospective study used one selected static US image per nodule, which may not fully reflect dynamic scanning or the full range of images available in routine practice. Second, the study was conducted at a single institution using a single AI platform, which may limit generalizability to other clinical environments or AI systems. Third, the sample contained a high proportion of malignant nodules (273/423, 64.5%), limiting the generalizability of predictive performance to populations with lower malignancy prevalence. Fourth, the same images were read first without and then with AI; the fixed reading order may have introduced recall or learning effects, without a formal washout interval. Fifth, because the reference standard included cytological diagnoses, potential misclassification due to false-positive or false-negative cytology cannot be entirely excluded (3,25). Sixth, the use of a dichotomous classification simplifies clinical decision-making and does not encompass the full complexity of risk stratification or biopsy recommendations. Therefore, future studies should focus on prospective, multicenter designs incorporating real-time US acquisition and broader clinical workflows to further clarify the role of AI assistance in thyroid nodule evaluation. Additional investigation is also warranted to determine how AI influences downstream clinical decisions, such as biopsy recommendations and patient management, particularly in resource-limited settings.

In conclusion, AI assistance did not significantly improve the diagnostic performance of physicians with intermediate experience in the US assessment of thyroid nodules. Interobserver agreement was numerically higher with AI, but the significance of this difference requires further evaluation.

Declarations

Author Contribution

B.B.S. contributed to Investigation and Writing–original draft. H.J.S., I.H.C., S.H., S.H.H., and H.J.K. contributed to Data curation. H.K. contributed to Writing–original draft. J.Y.K. contributed to Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Supervision, Funding acquisition, and Writing–review & editing. All authors reviewed and approved the final manuscript.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available because they contain patient-related information but may be available from the corresponding author on reasonable request, subject to institutional and ethical restrictions.

References

1. ADDIN ZOTERO_BIBL {"uncited":[],"omitted":[],"custom":[]} CSL_BIBLIOGRAPHY 1. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest. 2009 Aug;39(8):699–706. doi:10.1111/j.1365-2362.2009.02162.x
2. Tan GH, Gharib H. Thyroid Incidentalomas: Management Approaches to Nonpalpable Nodules Discovered Incidentally on Thyroid Imaging. Ann Intern Med. 1997 Feb 1;126(3):226–31. doi:10.7326/0003-4819-126-3-199702010-00009
3. 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 Jan;26(1):1–133. doi:10.1089/thy.2015.0020
4. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The Diagnosis and Management of Thyroid Nodules: A Review. JAMA. 2018 Mar 6;319(9):914. doi:10.1001/jama.2018.0898
5. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur Thyroid J. 2017;6(5):225–37. doi:10.1159/000478927
6. Shin JH, Baek JH, Chung J, Ha EJ, Kim J hoon, 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(3):370. doi:10.3348/kjr.2016.17.3.370

7. Ha EJ, Baek JH. Applications of machine learning and deep learning to thyroid imaging: where do we stand? Ultrasonography. 2021 Jan 1;40(1):23–9. doi:10.14366/usg.20068
8. 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 Sep;260(3):892–9. doi:10.1148/radiol.11110206
9. Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A, et al. An Ultrasonogram Reporting System for Thyroid Nodules Stratifying Cancer Risk for Clinical Management. J Clin Endocrinol Metab. 2009 May 1;94(5):1748–51. doi:10.1210/jc.2008-1724
10. Ha EJ, Chung SR, Na DG, Ahn HS, Chung J, Lee JY, et al. 2021 Korean Thyroid Imaging Reporting and Data System and Imaging-Based Management of Thyroid Nodules: Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol. 2021;22(12):2094. doi:10.3348/kjr.2021.0713
11. 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 May;14(5):587–95. doi:10.1016/j.jacr.2017.01.046
12. Grant EG, Tessler FN, Hoang JK, Langer JE, Beland MD, Berland LL, et al. Thyroid Ultrasound Reporting Lexicon: White Paper of the ACR Thyroid Imaging, Reporting and Data System (TIRADS) Committee. J Am Coll Radiol. 2015 Dec;12(12):1272–9. doi:10.1016/j.jacr.2015.07.011
13. Gitto S, Grassi G, De Angelis C, Monaco CG, Sdao S, Sardanelli F, et al. A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound. Radiol Med (Torino). 2019 Feb;124(2):118–25. doi:10.1007/s11547-018-0942-z
14. Hoang JK, Middleton WD, Farjat AE, Teefey SA, Abinanti N, Boschini FJ, et al. Interobserver Variability of Sonographic Features Used in the American College of Radiology Thyroid Imaging Reporting and Data System. Am J Roentgenol. 2018 Jul;211(1):162–7. doi:10.2214/AJR.17.19192
15. 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 Jun;247(3):762–70. doi:10.1148/radiol.2473070944
16. Wildman-Tobriner B, Taghi-Zadeh E, Mazurowski MA. Artificial Intelligence (AI) Tools for Thyroid Nodules on Ultrasound, From the *AJR* Special Series on AI Applications. Am J Roentgenol. 2022 Oct;219(4):547–54. doi:10.2214/AJR.22.27430
17. Xia M, Song F, Zhao Y, Xie Y, Wen Y, Zhou P. Ultrasonography-based radiomics and computer-aided diagnosis in thyroid nodule management: performance comparison and clinical strategy optimization. Front Endocrinol. 2023 May 12;14:1140816. doi:10.3389/fendo.2023.1140816
18. Ha EJ, Lee JH, Lee DH, Moon J, Lee H, Kim YN, et al. Artificial Intelligence Model Assisting Thyroid Nodule Diagnosis and Management: A Multicenter Diagnostic Study. J Clin Endocrinol Metab. 2024 Jan 18;109(2):527–35. doi:10.1210/clinem/dgad503
19. Xu D, Sui L, Zhang C, Xiong J, Wang VY, Zhou Y, et al. The clinical value of artificial intelligence in assisting junior radiologists in thyroid ultrasound: a multicenter prospective study from real clinical

practice. BMC Med. 2024 Jul 12;22(1):293. doi:10.1186/s12916-024-03510-z

20. Kang S, Lee E, Chung CW, Jang HN, Moon JH, Shin Y, et al. A beneficial role of computer-aided diagnosis system for less experienced physicians in the diagnosis of thyroid nodule on ultrasound. Sci Rep. 2021 Oct 14;11(1):20448. doi:10.1038/s41598-021-99983-6
21. Cece A, Agresti M, De Falco N, Sperlongano P, Moccia G, Luongo P, et al. Role of Artificial Intelligence in Thyroid Cancer Diagnosis. J Clin Med. 2025 Apr 2;14(7):2422. doi:10.3390/jcm14072422
22. Namsena P, Songsaeng D, Keatmanee C, Klabwong S, Kunapinun A, Soodchuen S, et al. Diagnostic performance of artificial intelligence in interpreting thyroid nodules on ultrasound images: a multicenter retrospective study. Quant Imaging Med Surg. 2024 May;14(5):3676–94. doi:10.21037/qims-23-1650
23. Koh J, Lee E, Han K, Kim EK, Son EJ, Sohn YM, et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. Sci Rep. 2020 Sep 17;10(1):15245. doi:10.1038/s41598-020-72270-6
24. Hertzberg BS, Kliewer MA, Bowie JD, Carroll BA, DeLong DH, Gray L, et al. Physician training requirements in sonography: how many cases are needed for competence? AJR Am J Roentgenol. 2000;174(5):1221–1227. doi:10.2214/ajr.174.5.1741221
25. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda System for Reporting Thyroid Cytopathology: a meta-analysis. Acta Cytol. 2012;56(4):333–339. doi:10.1159/000339959.

Figures

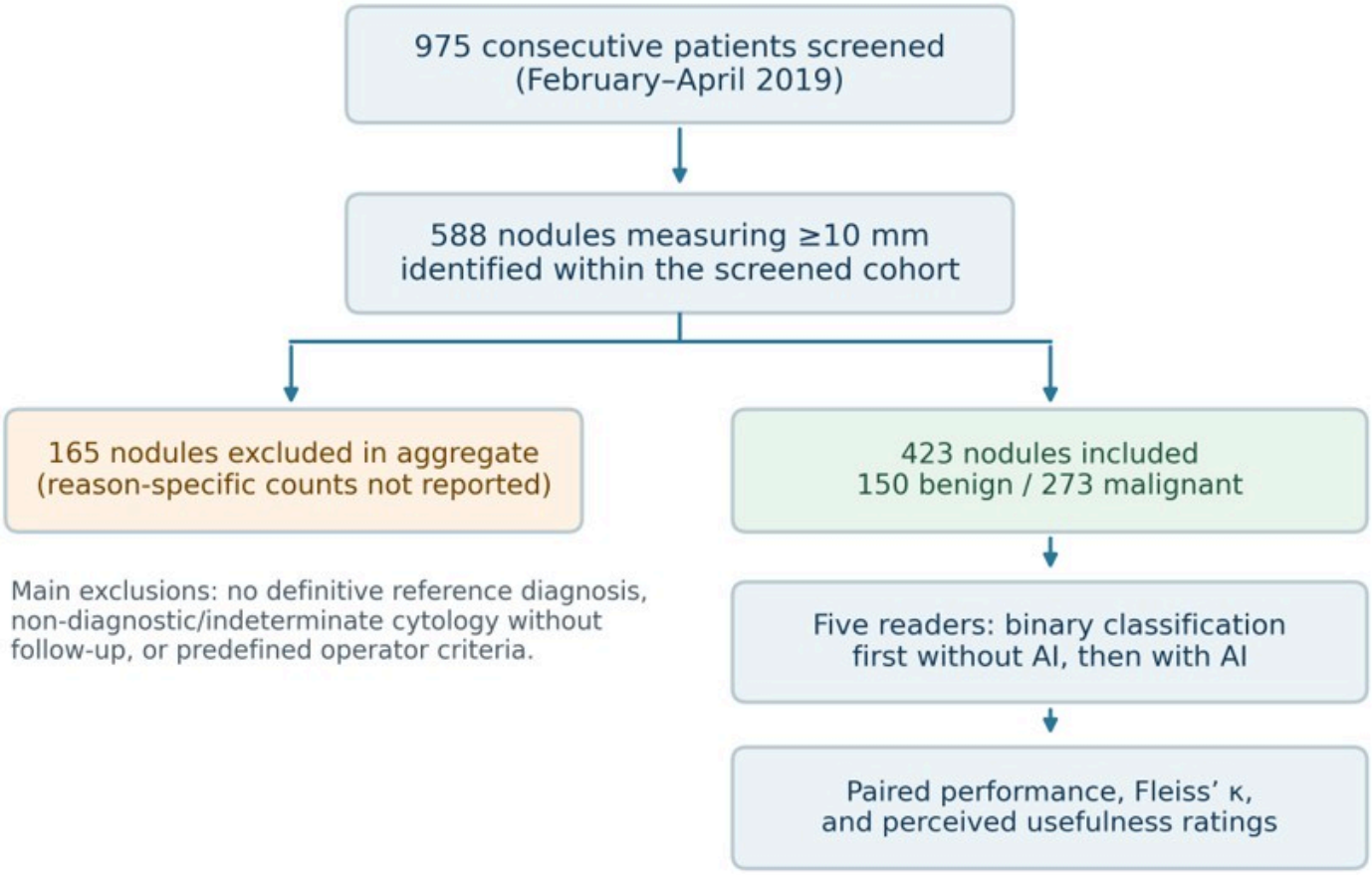


Figure 1

Study selection and reader-assessment flow.

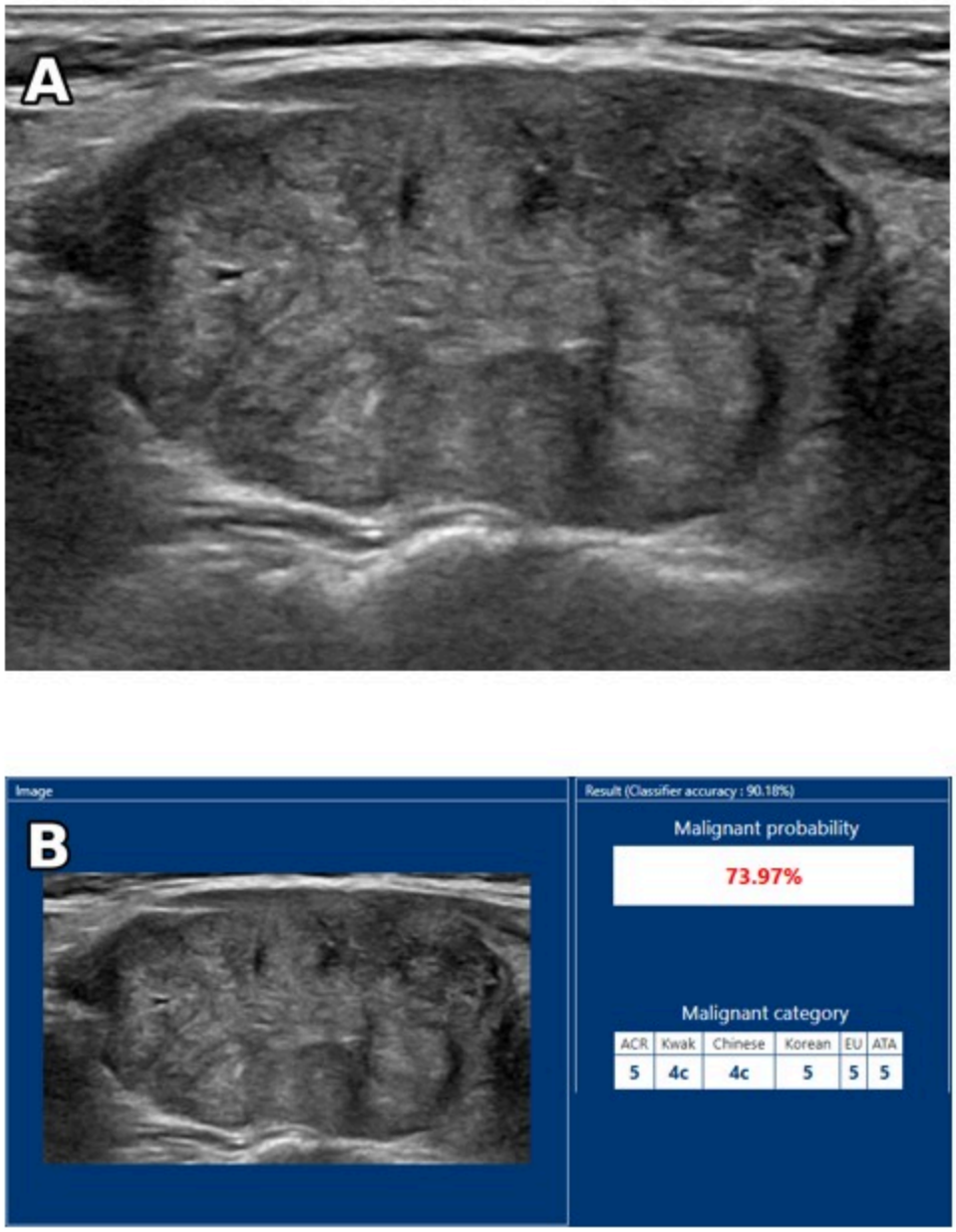


Figure 2

Representative case of a 44-mm thyroid nodule in a 66-year-old woman, confirmed as papillary thyroid carcinoma by fine-needle aspiration (A). Without AI assistance, three of the five readers initially classified the nodule as benign; all three revised their diagnosis to malignant after reviewing the AI output (B). The remaining two readers classified the nodule as malignant both with and without AI assistance.