THYROID
Volume 34, Number 4, 2024
© Mary Ann Liebert, Inc.
DOI: 10.1089/thy.2023.0603

Clinical Evaluation of an Artificial Intelligence-Based Decision Support System for the Diagnosis and American College of Radiology Thyroid Imaging Reporting and Data System Classification of Thyroid Nodules

Pablo Fernández Velasco,[1,2,*] Paloma Pérez López,[1,2] Beatriz Torres Torres,[1,2] Esther Delgado,[1,2]
Daniel de Luis,[1,2] and Gonzalo Díaz Soto[1,2,*]

Background: This study aimed to evaluate the clinical impact of an artificial intelligence (AI)-based decision support system (DSS), Koios DS, on the analysis of ultrasound imaging and suspicious characteristics for thyroid nodule risk stratification.

Methods: A retrospective ultrasound study was conducted on all thyroid nodules with histological findings from June 2021 to December 2022 in a thyroid nodule clinic. The diagnostic performance of ultrasound imaging was evaluated by six readers on the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) before and after the use of the AI-based DSS and by AI itself.

Results: A total of 172 patients (83.1% women) with a mean age of 52.3 ± 15.3 years were evaluated. The mean maximum nodular diameter was 2.9 ± 1.2 cm, with 11.0% being differentiated thyroid carcinomas. Among the nodules initially classified as ACR TI-RADS 3 and 4, AI reclassified 81.4% and 24.5% into lower risk categories, respectively. Receiver operating characteristic (ROC) curve analysis was performed to evaluate the diagnostic performance of the readers and the AI-based DSS versus histological diagnosis. There was an increase in the area under the ROC curve (AUROC) after the use of AI (0.776 vs. 0.817, $p < 0.001$). The AI-based DSS improved the mean sensitivity (Sens) (82.3% vs. 86.5%) and specificity (Spe) (38.3% vs. 54.8%), produced a high negative predictive value (94.5% vs. 96.4%), and increased the positive predictive value (PPV) (14.0% vs. 16.1%) and diagnostic precision (43.0% vs. 49.3%). Based on the ACR TI-RADS score, there was significant improvement in interobserver agreement after the use of AI ($r = 0.741$ for ultrasound imaging alone vs. 0.981 for ultrasound imaging and the AI-based DSS, $p < 0.001$).

Conclusions: The use of an AI-based DSS was associated with overall improvement in the diagnostic efficacy of ultrasound imaging, based on the AUROC, as well as an increase in Sens, Spe, negative and PPVs, and diagnostic accuracy. There was also a reduction in interobserver variability and an increase in the degree of concordance with the use of AI. AI reclassified more than half of the nodules with intermediate ACR TI-RADS scores into lower risk categories.

Keywords: thyroid nodule, AI, thyroid cancer, ultrasound

[1] Endocrinology and Nutrition Department, Hospital Clínico Universitario Valladolid, Valladolid, Spain.
[2] Centro de Investigación de Endocrinología y Nutrición Clínica (CIENC), Facultad de Medicina Universidad de Valladolid, Valladolid, Spain.

*These authors contributed equally to this work.

This work was presented as a scientific abstract at the 45th Annual Meeting of the European Thyroid Association in 2023 in Milan, Italy (September 9–12, 2023).

Introduction

The diagnosis of nodular thyroid pathology is becoming more frequent in clinical practice due to the generalization of imaging tests. Approximately 60% of randomly selected individuals have detectable thyroid nodules on ultrasound, especially women and older adults.[1,2] However, only 5% of these nodules are ultimately malignant.[3] There has been an increase in the incidence and prevalence of thyroid cancer diagnoses over the past decades; however, cancer-specific mortality has remained stable.[4]

Ultrasonography is the main imaging test to evaluate thyroid nodules; it represents the initial evaluation tool after physical examination. It allows to confirm the presence, number, and dimensions of nodules and to distinguish between those that should be analyzed by fine-needle aspiration (FNA) and those that can be followed by ultrasonography, according to their suspicious characteristics.[1,2]

However, the assessment of a thyroid nodule by ultrasound imaging has some drawbacks. Ultrasound imaging features suggestive of malignancy, such as hypoechogenicity, a mostly solid composition, a taller-than-wide shape, irregular margins, the absence of a halo, or the presence of intranodular calcification, are not specific enough to definitively diagnose malignancy on their own.[5] To solve this problem, malignancy risk stratification scales have been developed to integrate ultrasound information to standardize clinical decision-making.[6,7]

All risk stratification scales have demonstrated acceptable levels of sensitivity (Sens); however, the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) has the highest degree of specificity (Spe), reducing the number of unnecessary FNA biopsies and maintaining acceptable Sens.[6–8]

In this sense, decision support systems (DSSs) based on artificial intelligence (AI)/deep learning have recently been developed to assist clinicians in the interpretation of ultrasound imaging. These help to reduce the subjective component, thus decreasing inter- and intraobserver variability and improving the diagnostic performance of thyroid echography.[9,10] However, most of the results have focused on evaluating their diagnostic capacity in controlled studies and not in real clinical practice settings.[11]

Moreover, very few studies have evaluated the impact of DSS use on the diagnostic performance of each observer in a clinical setting with and without the support of an AI system, as well as that of the AI system alone.[12,13] Koios DS (Koios Medical, New York, NY) prepopulates existing ACR TI-RADS descriptors and provides a novel AI-derived risk assessment as an additional ACR TI-RADS descriptor (Koios AI Adapter) without changing any other elements of the existing TI-RADS lexicon.

Only one multicenter, cross-sectional, multireader validation study has been published.[14] However, this study did not reflect the true clinical activity in a thyroid nodule clinic, and the authors did not evaluate the impact of the AI Adapter.

The aim of the present study was to evaluate the impact of an AI-based DSS, Koios DS, on ultrasound imaging analysis and risk stratification based on ACR TI-RADS categorization in a real cohort of patients with nodular thyroid pathology evaluated in real-world practice.

Methods

This was a retrospective study of the ultrasound imaging of all nodules with cytological and/or histological results from a thyroid nodule clinic referral unit of a university hospital. All consecutive patients over 18 years of age with thyroid nodules and at least two unobstructed ultrasound images with cytologic and/or histologic findings evaluated from June 2021 to December 2022 were included. Patients with poor quality ultrasound images (image blur or nonstandard image acquisition), with incomplete cytologic and/or histologic data, or who refused to sign the informed consent form were excluded.

The following biochemical and clinical data were collected: sex, age at diagnosis, diagnostic method, relevant personal and family history, thyrotropin (TSH) and free thyroxine (fT4) levels, nodular size, percentage of malignancy determined by an FNA or biopsy of the surgical piece, and orthogonal images of the nodules under study (transverse/longitudinal) in the DICOM format. The images were analyzed by six readers before and after the use of the AI-based DSS.

At the time of the study, all readers were board-certified practicing physicians with 5–20 years of experience in thyroid ultrasound and ACR TI-RADS thyroid nodule evaluation. Each reader was blinded to the FNA and histological results to ensure an unbiased assessment of the nodules based solely on ultrasound images.

Each observer initially received a 30-minute training session to understand the results of the AI program, and a test to demonstrate the correct use of the AI platform based on the evaluation of five supervised test cases. All nodular images were presented to the observer in two orthogonal projections without delimiting the original regions of interest of the thyroid nodule. Each reader analyzed and recorded the composition, echogenicity, shape, margins, and echogenic foci of each nodule.

The reader assigned each characteristic a score and a risk category according to the criteria defined in the ACR TI-RADS risk assessment scale before and after the use of AI sequentially. That is, each reader scored and recorded the case twice, first as a preassessment by the unassisted observer (ultrasound [US]) and then as an AI-assisted reading with the assigned ACR TI-RADS features as well as an optional risk modifier generated by the AI-based DSS (US+AI), called the AI Adapter.

All thyroid nodule features and ACR TI-RADS risk classifications (US or US+AI) were mandatory, and the readers had the ability to edit all AI-generated features (including AI Adapter) during the US+AI condition. The order of the reading condition was randomized, and reading blocks were separated by a 2-week period.

All original images were acquired during clinical practice using GE LOGIC e7 ultrasound imaging equipment (Milwaukee, WI) by two board-certified physicians with 20 years of experience in thyroid nodule imaging. The internal scanning protocol included at least two orthogonal images for each nodule to capture different aspects of the nodule's morphology and characteristics with the best possible resolution. One of the board-certified physicians chose the most significant images (transverse and longitudinal) for review or excluded them based on the imaging quality. The selected images were presented in the transverse and longitudinal planes to each reader to comprehensively assess the nodules.

All malignant lesions were confirmed by post-thyroidectomy biopsy. Benign lesions were confirmed by post-thyroidectomy biopsy, if available, or by the FNA result based on Bethesda Il categorization. For nodules with a previous Bethesda II result, but with intermediate or high suspicious features, FNA was repeated to avoid false-negative cytology results, as recommended by current clinical guidelines.[1] Similarly, patients with indeterminate FNA classifications (Bethesda III and IV) were categorized as benign by repeat FNA or postsurgical biopsy, respectively.[1]

Finally, a total of 172 patients with thyroid nodules, comprising 172 nodules (11.0% biopsy-confirmed malignant nodules), were included in the study (Fig. 1). The cohort consisted of 83.1% female patients, with a mean ± standard deviation (SD) age of 55.3 ± 15.3 years. The nodules had a mean maximum transverse diameter of 2.9 ± 1.2 cm and median volume of 6.4 mL, with an interquartile range (IQR) of 1.6–13.8. The mean TSH level at diagnosis was 2.2 ± 1.9 mIU/L.

Additionally, 5.8% of the patients had a family history of thyroid cancer, and 18.6% were receiving thyroid hormone replacement therapy with levothyroxine. The baseline clinical characteristics are summarized in Table 1.

AI-based DSS

The AI system employed in this study uses computer vision and machine learning techniques to generate an engine capable of analyzing and interpreting the ultrasound image of the thyroid nodule. The characteristics of thyroid nodules categorized by the AI-based DSS used in the present work coincide exactly with the characteristics of suspicion defined in the ACR TI-RADS guidelines (with the exception of extrathyroidal extension),[7] as described previously by Barinov et al.[14]

The AI-based DSS evaluates the user-defined region of interest (ROI) of an ultrasound image corresponding to the thyroid nodule under study. From this image, the AI-based DSS categorizes (with a probability) each of the ACR TI-RADS components (composition, echogenicity, shape, margins, and echogenic foci). It also generates the AI Adapter, which is independent of the ACR TI-RADS classification and assessment. The AI Adapter allows for optional modification of the total risk by subtracting or adding to the total ACR TI-RADS score, specifically −2, −1, 0, +1, or +2 points (Supplementary Fig. S1).

Thus, the AI Adapter allows for incorporation of an independent machine learning-based thyroid nodule assessment to improve the ACR TI-RADS categorization beyond the assessment of each of the individual descriptors of the AI-based DSS itself.

Finally, based on the user's final total score (including the AI Adapter), the system makes a clinical action recommendation (essentially whether an FNA should be performed) for a specific thyroid nodule, following the same point and size thresholds of the ACR TI-RADS guidelines.[7]

Statistical analysis

The present study met the statistical requirements for sample size for an external validation diagnostic test accuracy study based on the area under the receiver operating characteristic curve (AUROC) for independent assessments

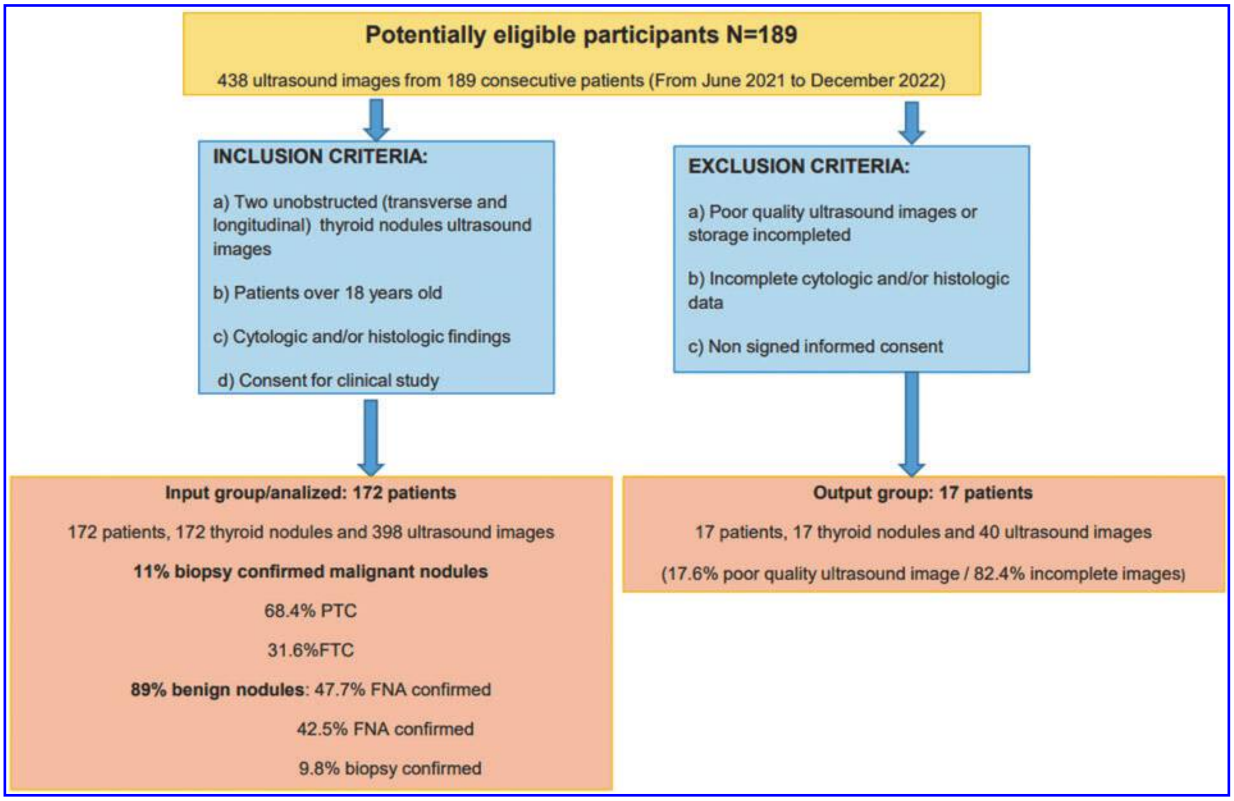
Potentially eligible participants N=189
438 ultrasound images from 189 consecutive patients (From June 2021 to December 2022)
INCLUSION CRITERIA:
a) Two unobstructed (transverse and longitudinal) thyroid nodules ultrasound images
b) Patients over 18 years old
c) Cytologic and/or histologic findings
d) Consent for clinical study
EXCLUSION CRITERIA:
a) Poor quality ultrasound images or storage incompleted
b) Incomplete cytologic and/or histologic data
c) Non signed informed consent
Input group/analized: 172 patients
172 patients, 172 thyroid nodules and 398 ultrasound images
11% biopsy confirmed malignant nodules
68.4% PTC
31.6%FTC
89% benign nodules: 47.7% FNA confirmed
42.5% FNA confirmed
9.8% biopsy confirmed
Output group: 17 patients
17 patients, 17 thyroid nodules and 40 ultrasound images
(17.6% poor quality ultrasound image / 82.4% incomplete images)

FIG. 1. Flowchart of the inclusion criteria and exclusion criteria for data collection. FNA, fine-needle aspiration; FTC, follicular thyroid cancer; PTC, papillary thyroid cancer.

Table 1. Baseline Characteristics of the Patients

Patients	n = *172*
Sex (female)	83.1%
Age	52.3 ± 15.3 years
Familiar thyroid cancer	5.8%
Personal cancer history	6.4%
Diagnostic approach	Physical examination 42.4% Computed tomography 12.3% Ultrasonography 31.6% Compressive symptoms 8.2% Functional disturbance 3.5%
Maximum nodular size	2.9 ± 1.2 cm
Transversal diameter	2.4 ± 1.1 cm
Anteroposterior diameter	1.9 ± 0.8 cm
Longitudinal diameter	2.9 ± 0.8 cm
Volume	6.4 mL [1.6–13.8]*
TSH	2.2 ± 1.9 mUI/L
fT4	1.3 ± 0.4 ng/dL
Levothyroxine treatment	18.6%
Positive autoimmunity	26.7%
Malignancy	11.0%
Papillary thyroid carcinoma	68.4%
Subtypes	
Classic	53.8%
Follicular	30.8%
Aggressive	15.4%
Follicular thyroid carcinoma	31.6%
Oncocytic variant	66.7%
Other variants	33.3%
Tumor staging AJCC	
Stage I	42.1%
Stage II	21.1%
Stage III	26.3%
Stage IV	10.5%

Data are presented as mean ± SD or median [IQR].

AJCC, American Joint Committee Cancer; fT4, free thyroxine; IQR, interquartile range; SD, standard deviation; TSH, thyrotropin.

on US and US+AI. The sample size was calculated to detect AUC (US and US+AI) >0.725, with a statistical power of 80%, significance level of 5%, and ratio of malignancy of 11% ($n = 149$). The Kolmogorov–Smirnov test was used to determine whether the variables followed a normal distribution.

Quantitative variables with a normal distribution are presented as mean ± SD, while quantitative variables with a non-normal distribution are presented as median [IQR]. Quantitative variables with a normal distribution were analyzed with Student's *t*-test. Nonparametric variables were evaluated using the Friedman and Wilcoxon tests. Qualitative variables are expressed as percentage (%) and were analyzed with the chi-square test or Fisher's exact test (when necessary).

For all analyses, an interpretation leading to a recommendation for FNA was considered a positive result for both US and US+AI.[14] The AUROC analysis was performed to determine diagnostic accuracy based on the ACR TI-RADS score by six readers before and after the use of an AI-based DSS. This analysis involved comparing the AUROCs and standard errors for 95% confidence intervals (CIs), with the histologic or cytologic result as the gold standard.

The Sens, Spe, positive predictive value (PPV), negative predictive value (NPV), and accuracy were assessed with a bilateral Z-test ($\alpha = 0.05$). All ratios were calculated based on the threshold for an FNA recommendation by using the ACR TI-RADS total score and thyroid nodule size criteria, according to the ACR TI-RADS guidelines.[7]

For all AUROC assessments and the ultrasound diagnostic accuracy assessment ratio, the absolute and relative differences are expressed as differences between AI-assisted assessment and assessment without the DSS (US+AI vs. US). Thus, positive values imply an improvement in the metric, which would support the use of AI, while negative values imply worse performance after the use of AI. The performance of the AI system alone (AI ACR TI-RADS features+AI Adapter), without reader intervention, was also analyzed.

Finally, to assess interobserver variability, Pearson correlation coefficients were calculated for the total ACR TI-RADS score of each observer averaged before and after the use of AI. A p-value <0.05 was considered to be statistically significant. SPSS Statistics, version 26 (IBM Corp, Armonk, NY), and RStudio (RStudio, Boston, MA) were used for the analysis. The study was approved by the Clinical Research Ethics Committee (CEIC) of the Hospital Center (PI 21-2525). All patients signed an informed consent form.

Results

The AUROC analysis was performed to determine the diagnostic accuracy of ultrasound imaging using the ACR TI-RADS scores determined by six readers and the AI-based DSS. The AI-DSS significantly improved the AUROC from 0.776 (CI 0.646–0.905) to 0.817 (CI 0.697–0.936) ($p < 0.001$). Overall, there was a mean increase of 5.3% (CI 3.4–7.9%) (Table 2) and all readers improved their AUROC (Fig. 2A, B).

Table 2. Impact of Artificial Intelligence-Based Decision Support Systems on Reader Performance, as Measured by the Area Under the Receiver Operating Characteristic Curve

Readers	*Difference in AUC* $(AUC_{US+AI}–AUC_{US})$ *[CI]*	*Percent change [CI]*
R1	0.040 [0.024 to 0.071]	5.330 [3.53 to 7.69]
R2	0.016 [–0.008 to 0.031]	2.097 [–1.433 to 5.556]
R3	0.036 [0.026 to 0.046]	4.774 [2.359 to 8.414]
R4	0.147 [0.055 to 0.203]	22.546 [13.861 to 36.895]
R5	0.028 [–0.006 to 0.050]	3.548 [2.295 to 4.223]
R6	0.009 [–0.012 to 0.023]	1.267 [0.530 to 1.637]
AI DSS alone	0.039 [0.020 to 0.059]	5.611 [4.946 to6.774]
Average	0.041 [0.025 to 0.072]	5.284 [3.421 to 7.892]

AI, artificial intelligence; AI-based DSS, AI-based decision support system; AUROC, area under the receiver operating characteristic curve; CI, 95% confidence interval; US, ultrasound imaging evaluation without AI; US+AI, ultrasound imaging evaluation with the AI decision support system.

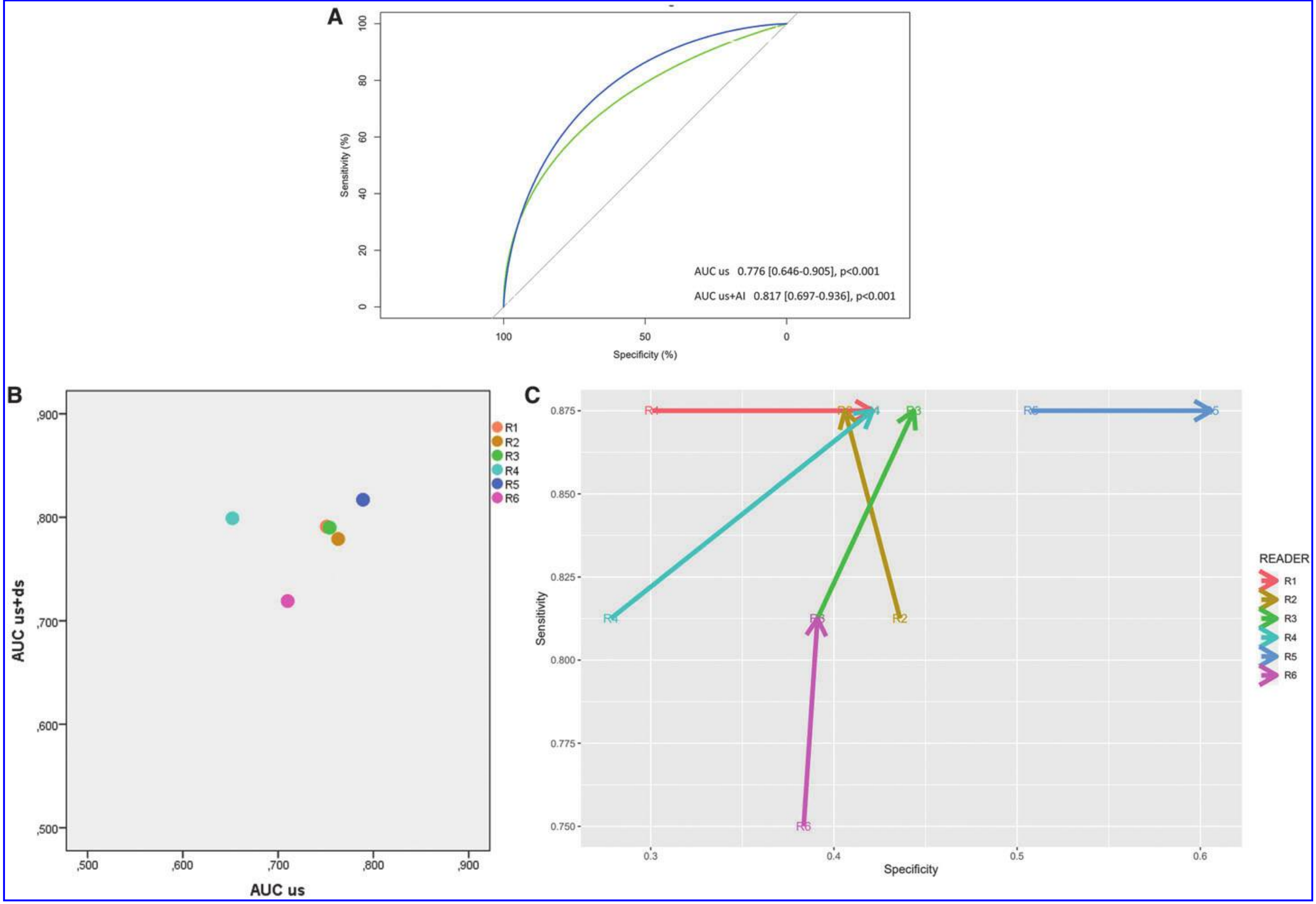

FIG. 2. (**A**) Average parametric ROC curve for the readers. (**B**) Comparison between AUCus alone and AUCus+DS individual reading conditions. (**C**) Comparison between the US and US+AI reading conditions regarding the change in FNA operating point per reader from US alone (base of arrow) to US+AI (head of arrow). AI, artificial intelligence; AUC, area under the ROC curve; ROC curve, receiver operating characteristic curve; us, ultrasound imaging evaluation without AI; us+AI, ultrasound imaging evaluation with the AI decision support system; Us+DS: ultrasound imaging evaluation with the AI decision support system.

Diagnostic accuracy, as assessed with Sens, Spe, NPV, PPV, and precision, was evaluated for the six readers as well as the AI system. The AI-based DSS improved Sens (from 82.29% to 86.46%), Spe (from 38.29% to 44.82%), NPV (from 94.53% to 96.39%), PPV (from 14.02% to 16.13%), and diagnostic accuracy (from 43.01% to 49.29%). The AI-based DSS showed a diagnostic accuracy similar to or slightly higher than that achieved by the observers with the use of AI (Sens = 81.25%, Spe = 53.03%, NPV = 95.89%, PPV = 13.83%, and accuracy = 56.08%) (Table 3 and Fig. 2C). Only one observer showed worse Spe, from 46.61% to 40.60% (−7.41%), after the use of AI.

There were no differences when assessing the accuracy of the AI system based on sex, family history of thyroid cancer, levothyroxine intake, circulating TSH levels, or the presence of autoimmunity or gland heterogeneity due to underlying thyroiditis.

Pearson's correlation analysis was used to evaluate interobserver variability in the total ACR TI-RADS score before and after the use of AI. There was significant improvement in interobserver variability after the use of AI ($r = 0.741$ for US and $r = 0.981$ for US+AI, $p < 0.001$) Figure 3.

Figure 4 shows how the AI-based DSS classified the analyzed nodules into ACR TI-RADS risk categories with and without the use of the AI Adapter. The analysis revealed that 82.8% and 24.5% of the nodules initially classified by AI as ACR TI-RADS 3 and 4, respectively, were reclassified into lower risk categories with the AI Adapter ($p < 0.001$). As a result, the AI Adapter eliminated the need for an FNA in 100% and 53.8% of those ACR TI-RADS 3 and 4 nodules, respectively, reclassified into lower risk categories (Supplementary Table S1).

Additionally, 11% of the analyzed nodules were initially categorized as ACR TI-RADS 1 or 2. This percentage increased to 42% of the total number of nodules evaluated with the AI Adapter ($p < 0.001$). The number of nodules classified as ACR TI-RADS 5 remained stable.

Discussion

The present study has demonstrated the usefulness of an AI-based DSS: it improved the ability of readers to discriminate malignant thyroid nodules (AUROC, Sens, Spe, PPV, NPV, and diagnostic accuracy), reduced interobserver variability, and increased the degree of agreement with the AI system. In addition, the AI-based DSS demonstrated similar and even slightly better diagnostic performance than the readers with previous experience in thyroid ultrasound and reclassified a significant percentage of nodules into lower risk categories, demonstrating its potential impact on clinical decision-making.

These findings highlight the potential of an AI-based DSS to enhance the diagnostic performance of ultrasound imaging in defining the risk of malignancy of thyroid nodules analyzed in a real cohort of patients with thyroid nodules evaluated in clinical practice. Moreover, the improved AUROC (5.3%) is similar to that found in other studies analyzing different AI-based systems (Table 2 and Fig. 2A).[9,10,15,16] Furthermore, the results are comparable with the only study published to date with the same AI system.[14]

However, most of the studies to date have only reported results based on image analysis, which do not correspond

Table 3. Impact of Artificial Intelligence-Based Decision Support Systems on Reader Performance, as Measured by Sensitivity, Specificity, Positive and Negative Predictive Values, and Diagnostic Accuracy

	Sens			*Spe*			*PPV*			*NPV*			*Accuracy*		
Reader	Sen_{US}	Sen_{US+AI}	*Percent change*	Spe_{US}	Spe_{US+AI}	*Percent change*	PPV_{US}	PPV_{US+AI}	*Percent change*	NPV_{US}	NPV_{US+AI}	*Percent change*	$Accuracy_{US}$	$Accuracy_{US+AI}$	*Percent change*
R1	87.50 [61.65–98.45]	87.50 [61.65–98.45]	0%	30.08 [22.43–38.63]	42.11 [33.60–50.97]	39.99%	13.08 [10.82–15.74]	15.38 [12.57–18.70]	17.58%	95.24 [84.21–98.68]	96.55 [88.29–99.05]	1.38%	36.24 [28.53–44.51]	46.98 [38.76–55.32]	29.63%
R2	81.25 [54.35–95.95]	87.50 [61.65–98.45]	7.69%	43.61 [35.03–52.47]	40.60 [32.18–49.46]	−7.41%	14.77 [11.59 18.64]	15.05 [12.32–18.27]	1.90%	95.08 [87.25 98.20]	96.43 [87.90–99.01]	1.42%	47.65 [39.41–55.98]	45.64 [37.46–53.99]	4.22%
R3	81.25 [54.35–95.95]	87.50 [61.65–98.45]	7.69%	39.10 [30.76–47.93]	44.36 [35.75.53.22]	13.45%	13.83 [10.90–17.40]	15.91 [12.96–19.38]	15.04%	94.55 [85.95–98.01]	96.72 [88.84–99.09]	2.30%	43.62 [35.53–51.98]	48.99 [40.72–57.30]	12.31%
R4	81.25 [54.35–95.95]	87.50 [61.65–98.45]	7.69%	27.82 [20.40–36.25]	42.11 [33.60–50.97]	51.37%	11.93 [9.47–14.91]	15.38 [12.57–18.70]	28.92%	92.50 [81.09–97.26]	96.55 [88.29–99.05]	4.38%	33.56 [26.04–41.74]	46.98 [38.76–55.32]	39.99%
R5	87.50 [61.65–98.45]	87.50 [61.65–98.45]	0%	50.76 [41.92–59.56]	60.61 [51.73–68.99]	19.41%	17.72 [14.32–21.73]	21.21 [16.89–26.29]	19.70%	97.10 [90.06–99.20]	97.56 [91.75–99.33]	0.47%	54.73 [46.35–62.92]	63.51 [55.21–71.26]	16.04%
R6	75.00 [47.62–92.73]	81.25 [54.35–95.95]	8.33%	38.35 [30.05–47.17]	39.10 [30.76–47.93]	1.96%	12.77 [9.67–16.68]	13.83 [10.90–17.40]	8.30%	92.73 [84.16–96.84]	94.55 [85.95–98.01]	1.96%	42.28 [34.24–50.64]	43.62 [35.53–51.98]	3.17%
AI DSS alone	81.25 [54.35–95.95]		—	53.03 [44.15–61.77]		—	17.33 [13.48–22.01]		—	95.89 [89.26–98.50]		—	56.08 [47.69–64.22]		—
Average	82.29%	86.46%	5.07%	38.28%	44.82%	17.08%	14.02%	16.13%	15.05%	94.53%	96.39%	1.97%	43.01%	49.29%	14.60%

NPV, negative predictive value; PPV, positive predictive value; Sen, sensitivity; Spe, specificity.

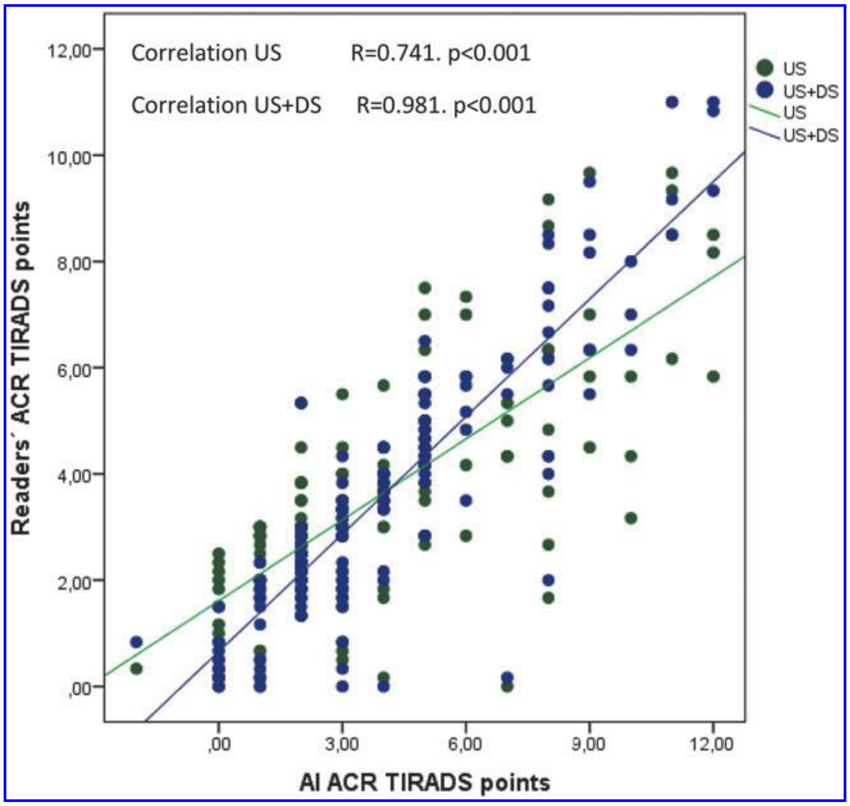

FIG. 3. Correlation between the readers and the AI ACR TI-RADS score. ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System.

to the daily practice of a thyroid nodule clinic, even with malignancy percentages well above the usual (20–70%).[14,17] Obviously, this high malignancy rate conditions the high PPV and low NPV of the test and may cause loss of diagnostic profitability of the AI system in a clinical practice setting.[18] However, the present study has demonstrated the usefulness of an AI-based DSS in a real imaging cohort corresponding to the clinical activity with cytology/histology over 18 months for a thyroid nodule reference unit in a population of >250,000 inhabitants.

In this context, it is necessary to highlight the very high NPV (96.39%), even with malignancy rates of 11% in the cohort (higher than expected due to the risk of malignancy of thyroid nodular pathology) (Table 3).[1] That is, the AI-based DSS could rule out malignancy in virtually all nodules, in which the AI system rejected the need for FNA. On the other hand, it is important to underline how the AI system improved the diagnostic yield and AUROC of all observers despite starting from high values of diagnostic capability without the use of AI (Fig. 2A–C).

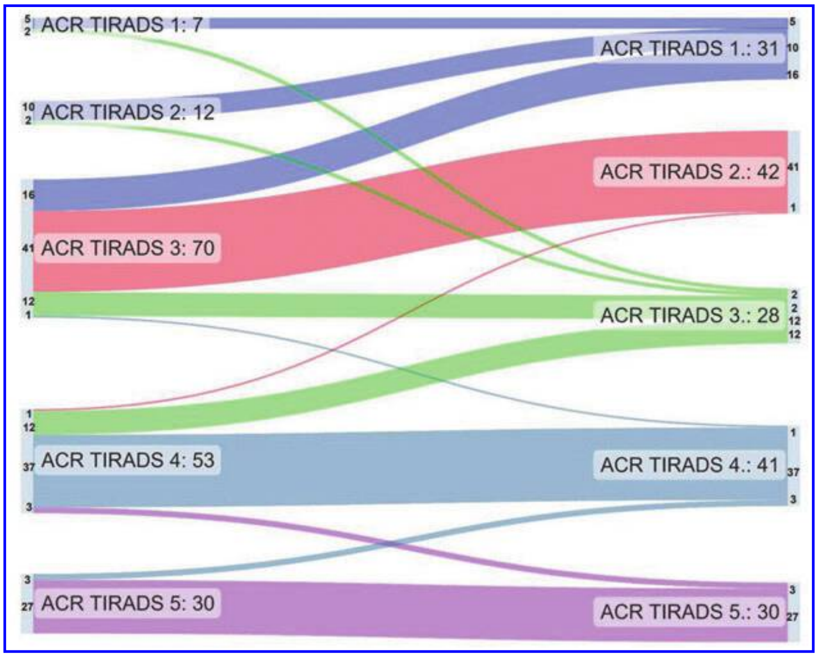

FIG. 4. ACR TI-RADS before and after the use of the Koios AI Adapter. DSS, decision support system without the AI Adapter; US+AI Adapter, decision support system with the AI Adapter.

Only one observer (R2) showed a reduction in Spe, but this change was accompanied by an increase in the AUROC similar to that of the five other readers. Similar results have been reported previously, and this highlights the occasional disconnect between point-based risk estimation (AUROC) and rule-based management pathways (FNA operating point), with the latter having a direct clinically relevant impact on patient care.[14]

Furthermore, the diagnostic performance of the AI-based DSS was similar to or slightly better than the board-certified, practicing, and highly experienced readers. It is possible that users with less experience in thyroid nodule evaluation may derive more benefit from the use of an AI-based DSS, as previous studies have shown, even those without specific prior knowledge.[17] This subgroup should definitely be analyzed in future studies.

The great increase in the incidence of thyroid nodule diagnoses due to the widespread availability and use of high-resolution ultrasonography poses a challenge in the diagnosis of thyroid cancer.[19] The aim of clinical guidelines and stratification scales (and therefore of AI for the analysis of thyroid nodules) should be to limit the number of FNA biopsies to those nodules in which ultrasound features are suggestive of malignancy, without reducing Sens.

This approach aims to avoid the health care and economic overload or the procedure and subsequent follow-up, as well as iatrogenesis and patient stress.[8] In this regard, analysis of the use of the AI Adapter for risk categorization of thyroid nodules using the current Koios DS is of special interest. In our study, the AI Adapter reclassified 70 nodules initially classified as ACR TI-RADS 3 or 4 into lower risk categories. This reclassification applied to 41% of the total nodules analyzed.

In total, 58 nodules were categorized as very low risk of malignancy and therefore did not require FNA, representing 33% of the total number of nodules analyzed (Fig. 4). Thereafter, nodules in those categories of lower risk and without major suspicion criteria[5] were reclassified by AI as benign, avoiding invasive and costly procedures. These AI-modified categories represent the lower risk nodules in which nodular size is the main criterion for FNA.

It is possible that future clinical guidelines will modify the size cutoffs or subdivide ACR TI-RADS categories 3 and 4 because these categories involve the greatest number of benign FNA biopsies and therefore the greatest health care and economic burden. On the other hand, the number of nodules classified as ACR TI-RADS 5 remained unchanged with the use of the AI Adapter, underscoring the need to avoid reductions in Sen (malignant nodules for which AI does not advise FNA) in an AI-based DSS.

The present study has certain limitations. First, the number of nodules and readers was relatively low compared with multicenter studies, and this study was retrospective.

Second, the analysis of thyroid nodules was restricted to the two most significant static orthogonal (transverse and longitudinal) images for each nodule with cytologic or biopsy results after thyroidectomy during the 18 months of the study. While it is noteworthy that the majority of current AI thyroid nodule software approaches rely on static images, this approach may potentially limit the diagnostic capacity of both the observer and the AI DSS.

Finally, malignancy diagnosis was restricted to differentiated (papillary or follicular) thyroid carcinoma.

This study also has notable strengths. The images and nodules analyzed correspond to a real imaging cohort with cytology/histology from an experienced thyroid nodule clinic that performs all cytological and ultrasound studies in its reference area. Moreover, the malignancy ratio is representative of the clinical reality.[11] All suspicious nodules underwent cytologic confirmation through at least two separate FNA biopsies or direct biopsy by thyroidectomy, especially those initially labeled as malignant, in accordance with current guidelines.[1]

This approach ensured the comparability of AI results with histologic results as the true gold standard. The present study analyzed the usefulness of AI in nodules with high or intermediate risk. The restriction of this study to those nodules with histological findings ensured the true classification of thyroid nodules as malignant or benign and restricted the use of AI to truly relevant clinical situations. Hence, this study avoided AI overanalysis of thyroid nodules with little clinical relevance (simple cysts and infracentimetric nodules, among others).[1,2]

On the other hand, the AI-based DSS required manual selection of an ROI by each reader, which could introduce bias and affect the reproducibility of AI results due to its subjective nature. There was an attempt to reduce this variability by implementing a training session and standardization of ROI selection for all readers. To date, all published Koios DS studies have used prespecified ROIs. However, this situation does not correspond to the actual DSS workflow.

Finally, the previously published studies did not collect or consider clinical data of major importance in the management of thyroid nodules or their influence on AI performance, namely thyroid function, the presence of biochemical autoimmunity or glandular heterogeneity due to thyroiditis, a personal and family history of thyroid cancer or cervical radiation, and manual ROI selection. In the present study, none of these variables influenced the usefulness of the AI-based DSS.

In conclusion, the use of an AI-based DSS was associated with an overall improvement in the diagnostic capability of ultrasound imaging measured by the AUROC, as well as an increase in the Sens, Spe, NPV, PPV, and diagnostic accuracy of readers. There was also a reduction in interobserver variability and an increase in the degree of concordance with the use of AI. AI reclassified more than half of the nodules with intermediate ACR TI-RADS scores into lower risk categories.

Authors' Contributions

P.F.V. was involved in investigation and writing—original draft; P.P.L., B.T.T., and E.D. were involved in investigation; D.d.L. was involved in funding acquisition; and G.D.S. was involved in funding acquisition, supervision, and writing—review and editing.

Author Disclosure Statement

Koios DS was provided free of charge by Koios Medical, Inc., for this study. The design, execution, data collection, and analysis of the study were performed independently by the research investigators. The software company was not involved in any way in the design, execution, data collection, or analysis of the study, or any other phase, or in any financial relationship with the researchers.

Funding Information

This research received a research grant from Sociedad Castellano y Leonesa de Endocrinología, Diabetes y Nutrition (Scledyn) 2022.

Supplementary Material

Supplementary Table S1
Supplementary Figure S1

References

1. 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; doi: 10.1089/thy.2015.0020
2. 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; doi: 10.1530/ETJ-23-0067
3. Durante C, Grani G, Lamartina L, et al. The diagnosis and management of thyroid nodules: A review. JAMA 2018; 319(9):914–924; doi: 10.1001/jama.2018.0898
4. Lim H, Devesa SS, Sosa JA, et al. Trends in thyroid cancer incidence and mortality in the United States, 1974–2013. JAMA 2017;317(13):1338–1348; doi: 10.1001/jama.2017.2719
5. Rago T, Vitti P. Risk stratification of thyroid nodules: From ultrasound features to TIRADS. Cancers 2022;14(3):717; doi: 10.3390/cancers14030717
6. Middleton WD, Teefey SA, Reading CC, et al. Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association guidelines. Am J Roentgenol 2018;210(5):1148–1154; doi: 10.2214/AJR.17.18822
7. Tessler FN, Middleton WD, Grant EG, 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(5):587–595; doi: 10.1016/j.jacr.2017.01.046
8. Hoang JK, Middleton WD, Langer JE, et al. Comparison of thyroid risk categorization systems and fine-needle aspiration recommendations in a multi-institutional thyroid ultrasound registry. J Am Coll Radiol 2021;18(12):1605–1613; doi: 10.1016/j.jacr.2021.07.019
9. Jin Z, Zhu Y, Zhang S, et al. Diagnosis of thyroid cancer using a TI-RADS-based computer-aided diagnosis system: A multicenter retrospective study. Clin Imaging 2021;80: 43–49; doi: 10.1016/j.clinimag.2020.12.012
10. Zhu Y-C, Jin P-F, Bao J, et al. Thyroid ultrasound image classification using a convolutional neural network. Ann Transl Med 2021;9(20):1526; doi: 10.21037/atm-21-4328
11. Toro-Tobon D, Loor-Torres R, Duran M, et al. Artificial intelligence in thyroidology: A narrative review of the current applications, associated challenges, and future directions. Thyroid 2023;33(8):903–917; doi: 10.1089/thy.2023.0132
12. Ludwig M, Ludwig B, Mikuła A, et al. The use of artificial intelligence in the diagnosis and classification of thyroid nodules: An update. Cancers 2023;15(3):708; doi: 10.3390/cancers15030708
13. Tessler FN, Thomas J. Artificial intelligence for evaluation of thyroid nodules: A primer. Thyroid 2023;33(2):150–158; doi: 10.1089/thy.2022.0560
14. Barinov L, Jairaj A, Middleton WD, et al. Improving the efficacy of ACR TI-RADS through deep learning-based descriptor augmentation. J Digit Imaging 2023;36(6):2392–2401; doi: 10.1007/s10278-023-00884-z
15. Wu M-H, Chen K-Y, Shih S-R, et al. Multi-reader multicase study for performance evaluation of high-risk thyroid ultrasound with computer-aided detection. Cancers 2020; 12(2):373; doi: 10.3390/cancers12020373
16. Kim HL, Ha EJ, Han M. Real-world performance of computer-aided diagnosis system for thyroid nodules using ultrasonography. Ultrasound Med Biol 2019;45(10):2672–2678; doi: 10.1016/j.ultrasmedbio.2019.05.032
17. Li Y, Liu Y, Xiao J, et al. Clinical value of artificial intelligence in thyroid ultrasound: A prospective study from the real world. Eur Radiol 2023;33(7):4513–4523; doi: 10.1007/s00330-022-09378-y
18. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: Sensitivity, specificity, positive predictive value, and negative predictive value. Med Kaunas Lith 2021;57(5):503; doi: 10.3390/medicina57050503
19. Jegerlehner S, Bulliard J-L, Aujesky D, et al. Overdiagnosis and overtreatment of thyroid cancer: A population-based temporal trend study. PLoS One 2017;12(6):e0179387; doi: 10.1371/journal.pone.0179387

Address correspondence to:
Gonzalo Díaz Soto, MD, MSc, PhD
Servicio Endocrinología y Nutrición
Hospital Clínico Universitario
Avenida Ramón y Cajal, 3
Valladolid 47005
Spain

E-mail: diazsotogonzalo@gmail.com;
dluisro@saludcastillaleon.es

This article has been cited by:

1. Johnson Thomas. 2024. Improving Interobserver Agreement in Thyroid Nodule Evaluation: A Clinical Review of an AI-Based Decision Support System. *Clinical Thyroidology®* **36**:6, 231-233. [Citation] [Full Text] [PDF] [PDF Plus]