endocrines

MDPI

Article

Prospective Optimization of Malignancy Risk Prediction in Indeterminate Thyroid Nodules: Diagnostic Synergy of ACR TI-RADS and the 2023 Bethesda System

Ozlem Aydin [1], Bulent Colakoglu [2], Cavit Kerem Kayhan [1], Mehmet Güven Günver [3], Mariana Simplício [4], Joana Pinto Schmitt [5] and Sule Canberk [4,*]

[1] Department of Pathology, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul 34638, Turkey; ozlem.aydin@acibadem.edu.tr (O.A.); cavit.kerem.kayhan@acibadem.com (C.K.K.)
[2] Department of Radiology, American Hospital, Istanbul 34365, Turkey
[3] Department of Biostatistics, İstanbul Faculty of Medicine, İstanbul University, Istanbul 34093, Turkey; guven.gunver@istanbul.edu.tr
[4] Department of Pathology, RISE-Health, Faculty of Medicine of the University of Porto (FMUP), Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal; mariana.brito@ulssjoao.min-saude.pt
[5] Radiology Unit, ULS Trás-os-Montes e Alto Douro, 5000-51 Vila Real, Portugal; joanapintodx@gmail.com
* Correspondence: sulecanberk@med.up.pt

Abstract

Background: Risk stratification of indeterminate thyroid nodules (Bethesda III–IV) remains difficult and often triggers unnecessary procedures. Ultrasound-based ACR TI-RADS and the 2023 Bethesda System are widely used, but the incremental value of combining them and the role of size thresholds needs prospective validation. **Objective:** The objective of this study was to prospectively compare the diagnostic performance of ACR TI-RADS and the 2023 Bethesda System, alone and in combination, for predicting malignancy in thyroid nodules, with dedicated analyses of indeterminate lesions (Bethesda categories III–IV), including subtypes of Bethesda III (nuclear atypia vs. other atypia), and the impact of nodule size. **Methods:** Histopathology was available for 131 nodules. Diagnostic metrics (sensitivity, specificity, PPV, NPV), ROC curves (DeLong comparison), and Youden indices were calculated for individual and combined thresholds; a 16 mm size cut-off was explored. **Results:** Malignancy was confirmed in 105/131 nodules (80.2%). Bethesda outperformed TI-RADS (AUC 0.87 vs. 0.69; DeLong $p = 0.041$). Malignancy rates rose with higher categories (e.g., TI-RADS 5: 93.6%; Bethesda category V: 100%; Bethesda category VI: 100%) and were markedly elevated in the histologically confirmed subset for Bethesda category III (32/41; 78.0%) and IV (6/8; 75.0%). The combined requirement of TI-RADS ≥ 4 and Bethesda ≥ 4 maximized specificity (96.2%) and PPV (98.4%) with a high Youden J (0.552), supporting a rule-in strategy in category IV of Bethesda. Size alone was a weak discriminator (AUC 0.66); within Bethesda III–IV nodules, malignancy did not differ significantly by the 16 mm threshold ($p = 1.00$). ROC using continuous tumor size yielded AUC = 0.66; the ROC-derived optimal cut-off was 16 mm. Applying this split produced sensitivity 0.80 and specificity 0.50. **Conclusions:** Integrating ACR TI-RADS with Bethesda cytology significantly improves specificity and PPV for indeterminate thyroid nodules, supporting a morphology-driven approach over traditional size-based thresholds. Incorporation of combined sonographic–cytologic criteria into management algorithms may reduce unnecessary interventions and optimize patient care.

Keywords: thyroid nodule; Bethesda III–IV; ACR TI-RADS; ultrasound; cytology; malignancy risk; diagnostic accuracy; FNA; fine needle aspiration

Academic Editor: Antonio Brunetti

Received: 19 January 2026
Revised: 28 February 2026
Accepted: 11 March 2026
Published: 19 March 2026

1. Introduction

Accurate differentiation between malignant and benign thyroid nodules remains a central challenge in endocrine diagnostics, particularly when cytological evaluation yields indeterminate results. The increased clinical utilization of high-resolution ultrasonography (US) has significantly heightened thyroid nodule detection rates, creating a parallel imperative to refine diagnostic stratification and minimize overtreatment [1–3]. Currently, one of the most widely used risk-stratification tools is the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS); a meta-analysis reported pooled sensitivity 0.89, specificity 0.70, and AUROC 0.86 [4,5] and fine-needle aspiration (FNA) cytology was interpreted through the Bethesda System for Reporting Thyroid Cytopathology [6,7]. However, uncertainty persists, particularly for Bethesda category III (Atypia of Undetermined Significance, AUS) and Bethesda category IV (Follicular Neoplasm, FN) categories, often resulting in repeated aspirations, diagnostic surgeries, and increased healthcare costs [6–8].

The ACR TI-RADS employs a standardized sonographic lexicon, stratifying nodules by composition, echogenicity, margins, shape, and echogenic foci [4]. Concurrently, the 2023 Bethesda update subdivides the AUS category into nuclear (AUS-N) and other atypia (AUS-O) to refine Risk of Malignancy (ROM) estimates [6,7]. Reported ROM for AUS ranges from about 6% to 42% (and up to ~50% in larger cohorts), is higher in nuclear than architectural/FLUS atypia, and increases with higher TI-RADS strata [6,7,9–11].

Yet, despite these refinements, Bethesda III–IV nodules still present diagnostic ambiguity, motivating evaluation of combined stratification models [10,11]. Comparative analyses of ultrasound risk systems show no single universal standard, but ACR TI-RADS frequently ranks highest for both sensitivity and specificity at its optimal (TR5) threshold [12].

Several retrospective studies have suggested potential synergy when integrating ACR TI-RADS with Bethesda scoring; however, these have largely relied upon historical cytologic categories or lacked robust prospective validation [13,14]. Additionally, emerging evidence challenges traditional size-based biopsy thresholds, indicating that smaller nodules may harbor significant malignancy risk if coupled with concerning morphological or cytologic features [15–18]. This underlines a critical need to re-evaluate the clinical reliance on size thresholds in favor of morphology-driven decision pathways.

In response, we designed a prospective, single-center study involving 1009 thyroid nodules assessed through contemporaneous TI-RADS scoring and the 2023 Bethesda System, with histopathological confirmation available for 131 nodules. Our primary objective was to determine whether combining TI-RADS with Bethesda scores enhances malignancy prediction within the diagnostically ambiguous Bethesda categories III–IV. A secondary aim investigated the influence of nodule size in relation to malignancy risk within this dual-system framework. Lastly, as an exploratory analysis, we evaluated differences between the AUS subcategories (AUS-N vs. AUS-O) concerning their malignancy rates and sonographic profiles.

By systematically evaluating a large-scale, prospectively acquired dataset, this study aims to establish a robust, real-world diagnostic synergy between ACR TI-RADS and the updated Bethesda System. The findings promise to provide clear, evidence-based thresholds, minimizing unnecessary interventions and enhancing clinical decision-making accuracy in managing thyroid nodules.

2. Material and Method

2.1. Study Design and Setting

This prospective, single-center, cross-sectional cohort study was conducted between September 2019 and August 2024 at a tertiary academic hospital. All procedures were

conducted in accordance with institutional ethical standards and the Declaration of Helsinki [19]. The study protocol was approved by the Institutional Review Board of the institution in compliance with Good Clinical Practice guidelines (approval number: 2025-06/39) [20].

2.2. Study Population

We included patients undergoing US-guided FNA of thyroid nodules that were classified by both the ACR TI-RADS system and Bethesda system. We identified 1009 consecutive nodules with contemporaneous ACR TI-RADS scoring and cytology. Of these, 131 (13.0%) subsequently underwent surgery with histopathology available; diagnostic accuracy and ROM analyses using histology as reference standard were restricted to this surgically treated subset.

A total of 1009 consecutive patients were enrolled, with one index nodule per patient included in the analysis. Inclusion criteria were: (1) age $\geq$18 years, (2) availability of both cytology [6] and ACR TI-RADS scores, and (3) adequate FNA cellularity per Bethesda standards. Exclusion criteria included: (1) non-diagnostic cytology (Bethesda I), (2) purely cystic or anatomically normal nodules (TI-RADS 1), (3) noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP), well-differentiated tumors of uncertain malignant potential (WDT-UMP) and follicular tumor of uncertain malignant potential (FT-UMP) in pathology and were not counted as malignant in ROM calculations, and (4) patients with a history of neck irradiation. The number of cases with available molecular results was very small and therefore these data were not included in the statistical analyses.

2.3. Ultrasound Evaluation and Fine-Needle Aspiration and Cytopathology and Histopathology

All US examinations were performed by a single expert radiologist. The ACR TI-RADS 2017 [4] lexicon was applied prospectively using a ML6–15 or ML4–20 MHz high-frequency linear transducer. The radiologist was blinded to cytological diagnoses during image acquisition and TI-RADS scoring, ensuring unbiased assessment. All nodules were classified into TI-RADS categories 2–5 based on composition, echogenicity, margins, shape, and echogenic foci.

FNA was performed on the same day under US guidance using a 22-gauge needle (hypodermic needle—0.7 $\times$ 32 mm 22 g, Berika Teknoloji Medikal, Konya, Turkey) with local anesthesia with cold spray. Rapid on-site evaluation (ROSE) was conducted for each patient using an air-dried Diff-Quik stain to ensure adequacy and a May-Grünwald–Giemsa (MGG) stain was used as permanent stain. Cytologic interpretation was performed by a single expert cytopathologist, who was aware of US findings but blinded to histologic outcomes. The 2023 update of the Bethesda system [6] was used to classify all samples.

Of the 1009 nodules, 131 (13.0%) underwent definitive histopathologic evaluation via surgical resection (lobectomy or total thyroidectomy).

2.4. Data Analysis and Diagnostic Stratification

All statistical analyses were performed using IBM SPSS Statistics (Version 29.0.1.1) and Python (v3.11, Scikit-learn and Statsmodels libraries) for advanced modeling and visualization.

Descriptive statistics were calculated for all continuous and categorical variables. Categorical variables were compared using the Chi-square test or Fisher's exact test where appropriate. Continuous variables were compared using Student's *t*-test (for normally distributed data) or the Mann–Whitney U test (for non-normally distributed data), based on Kolmogorov–Smirnov tests of normality.

To evaluate diagnostic performance, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Youden index (defined as sensitivity + specificity − 1; higher values indicate better overall test performance) were calculated for individual systems (TI-RADS and Bethesda) and for selected combinations. ROC (Receiver Operating Characteristic) curves were constructed for TI-RADS, Bethesda category, and tumor size, using histopathological diagnosis as the reference standard, and compared using the DeLong test. AUROC (Area Under the ROC Curve) values were used to summarize diagnostic accuracy.

We summarized diagnostic accuracy using ROC AUC. For tumor size, ROC was run with size as a continuous predictor. The optimal size cut-off was identified by maximizing the Youden index (J = Se + Sp − 1) and then applied as a single threshold; performance for the 16 mm split (Se 0.80, Sp 0.50) is reported separately from the continuous-size AUC. This ROC-derived cut-off (16 mm) was explored as a data-driven, illustrative threshold rather than a guideline recommendation; ACR TI-RADS employs category-specific size triggers (e.g., TR4 $\geq$ 15 mm, TR5 $\geq$ 10 mm) for biopsy decisions, which remain the standard clinical framework.

All statistical tests were two-sided, and p-values <0.05 were considered statistically significant.

3. Results

Cohort Characteristics and Nodule Classification: A total of 1009 thyroid nodules were prospectively evaluated using both ACR TI-RADS and the 2023 Bethesda System. Patient age ranged from 18 to 85 years, with a median age of 47 years. Median nodule size was 14 mm (IQR: 10–21 mm). The majority of nodules fell into Bethesda II (66.1%) and TI-RADS 3 (46.6%) categories. Marked sonographic–cytologic mismatch was observed, most notably in 38 nodules classified as TI-RADS 5 despite Bethesda II cytology; an additional 27 nodules were classified as TI-RADS 5 with Bethesda III cytology (Table 1).

Table 1. Distribution of 1009 Thyroid Nodules by ACR TI-RADS Category and Bethesda Classification.

ACR TI-RADS (*n*)	Bethesda II	Bethesda III	Bethesda IV	Bethesda V	Bethesda VI	Total
ACR TI-RADS 2 (69)	63 (91.3%)	3 (4.3%)	1 (1.4%)	0 (0%)	2 (2.9%)	69 (6.8%)
ACR TI-RADS 3 (470)	374 (79.6%)	85 (18.1%)	6 (1.3%)	5 (1.1%)	0 (0%)	470 (46.6%)
ACR TI-RADS 4 (328)	192 (58.5%)	83 (25.3%)	21 (6.4%)	8 (2.4%)	24 (7.3%)	328 (32.5%)
ACR TI-RADS 5 (142)	38 (26.8%)	27 (19.0%)	1 (0.7%)	26 (18.3%)	50 (35.2%)	142 (14.1%)
Total (1009)	667 (66.1%)	198 (19.6%)	29 (2.9%)	39 (3.9%)	76 (7.5%)	1009 (100%)

Malignancy Rates and Diagnostic Concordance: Histopathological confirmation was available for 131 nodules, among which 105 (80.2%) were malignant. Malignancy rates increased significantly with higher Bethesda and TI-RADS categories. TI-RADS 5 nodules had a malignancy rate of 93.6%, while Bethesda category VI and category V nodules demonstrated malignancy rates of 100%, respectively. In the surgically resected subset, Bethesda categories III and IV had surgical malignancy rates of 78.0% (32/41) and 75.0% (6/8), respectively (Table 2). However, only 41 of 198 Bethesda III nodules underwent surgery; therefore, a conservative minimum malignancy rate across all Bethesda III nodules is 32/198 = 16.2% (assuming nonresected nodules are benign).

Table 2. Malignancy rates by Bethesda and TI-RADS category.

ROM (Counts (Total/ Malignant), % Malignant)	Bethesda					
TI-RADS	**II**	**III**	**IV**	**V**	**VI**	**Row Total**
2	2/0 (0.0%)	1/0 (0.0%)	–	–	–	3/0 (0.0%)
3	10/5 (50.0%)	17/15 (88.2%)	2/1 (50.0%)	2/2 (100.0%)	–	31/23 (74.2%)
4	8/3 (37.5%)	16/10 (62.5%)	6/5 (83.3%)	3/3 (100.0%)	17/17 (100.0%)	50/38 (76.0%)
5	3/0 (0.0%)	7/7 (100.0%)	–	13/13 (100.0%)	24/24 (100.0%)	47/44 (93.6%)
Column Total	23/8 (34.8%)	41/32 (78.0%)	8/6 (75.0%)	18/18 (100.0%)	41/41 (100.0%)	131/105 (80.2%)

Cells show Total/Malignant with % malignant in parentheses; computed on the histology subset (n = 131).

Diagnostic Accuracy: ROC Curve Analysis demonstrated superior performance of the Bethesda scoring system compared to TI-RADS, with respective AUC values of 0.87 and 0.69 (DeLong test, p = 0.041), as shown in the ROC curve analysis. Tumor size alone had limited discriminative ability with an AUC of 0.66, reinforcing the inadequacy of relying solely on size for malignancy prediction (Figure 1).

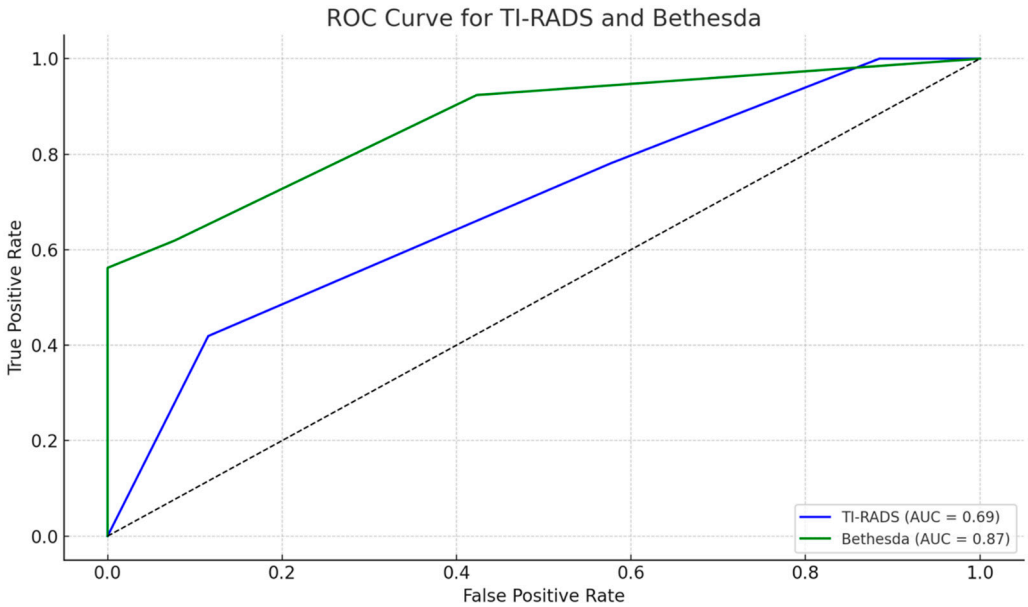

Figure 1. ROC curve comparing ACR TI-RADS and Bethesda 2023 classification in all histologically confirmed thyroid nodules. ROC curves comparing ACR TI-RADS and the 2023 Bethesda System in histologically confirmed nodules (n = 131). AUCs: TI-RADS = 0.69; Bethesda = 0.87. Curves are step functions at thresholds (TI-RADS $\geq$ 5, $\geq$4, $\geq$3; Bethesda $\geq$ 6, $\geq$5, $\geq$4, $\geq$3). The dashed diagonal represents random (chance) classification, corresponding to an AUC of 0.50. Any ROC curve that lies above this line demonstrates discriminatory power, whereas a curve on the line would add no diagnostic information.

Optimal Diagnostic Thresholds Evaluation of dual-threshold: We evaluated dual-threshold strategies combining TI-RADS and Bethesda (Figure 2; Supplementary Table S1). The combination of TI-RADS $\geq$ 4 and Bethesda $\geq$ 4 achieved Se 59.0%, Sp 96.2%, PPV 98.4%, NPV 36.8%, and a high Youden index (J = 0.552). Although TI-RADS $\geq$ 3 and Bethesda $\geq$ 5 yielded the numerically highest Youden J in the tested grid (0.562), this rule is largely driven by Bethesda V (which is already managed as high suspicion and mostly with a surgical approach), whereas TI-RADS $\geq$ 4 AND Bethesda $\geq$ 4 is clinically more informative for the gray-zone Bethesda III–IV categories where triage decisions are most challenging.

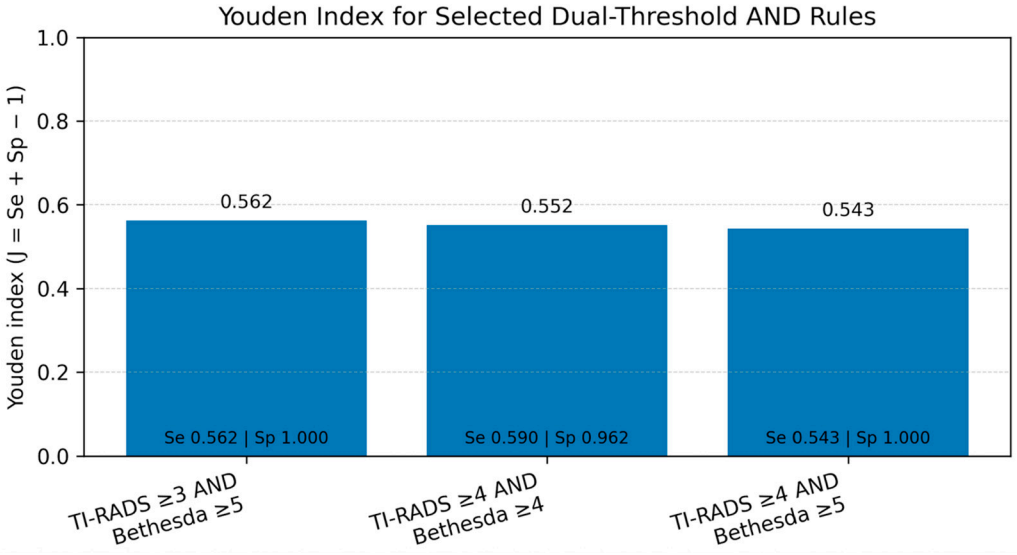

Figure 2. Youden Index for selected dual-threshold AND rules (TI-RADS + Bethesda). Youden index (J = Se + Sp − 1) for selected dual-threshold AND rules: TI-RADS ≥ 3 AND Bethesda ≥ 5, TI-RADS ≥ 4 AND Bethesda ≥ 4, and TI-RADS ≥ 4 AND Bethesda ≥ 5. Values are computed on the histology subset (n = 131). Sensitivity and specificity for each rule are shown beneath the bars.

In contrast, OR strategies increased sensitivity at the cost of specificity; for example, TI-RADS ≥ 4 OR Bethesda ≥ 3 yielded Se 95.2% and Sp 26.9% (J = 0.222), and TI-RADS ≥ 4 OR Bethesda ≥ 4 yielded Se 81.0% and Sp 38.5%. Some OR rules (e.g., 3 OR 3) showed very high NPV in this surgically enriched subset because the test-negative group was small; however, their very low specificity limits rule-out utility and generalizability. Overall, no combined strategy achieved both high sensitivity and high specificity; the selected AND rules function best as rule-in thresholds, whereas OR rules are not suitable as standalone rule-out criteria.

Nodule Size and Malignancy Risk: Among the 131 histologically confirmed nodules, those <16 mm unexpectedly showed a higher malignancy rate than larger nodules (86.6% vs. 61.8%; p = 0.004; OR 4.0, 95% CI 1.62–9.89; Table 3). This paradox likely reflects selective FNA of small nodules harboring suspicious sonographic or cytologic features. In the indeterminate (Bethesda categories III–IV) subgroup, however, nodule size had no impact: malignancy rates were virtually identical for <16 mm (77.1%) and ≥16 mm (78.6%; Fisher's p = 1.00; OR 0.92, 95% CI 0.21–4.13) (Supplementary Table S3). Consistent with this, ROC treating size as continuous gave AUC = 0.66, whereas the single 16 mm split achieved Se 0.80 and Sp 0.50.

Table 3. Malignancy rates and diagnostic accuracy for a 16 mm pathologic size cut-off (n = 131).

Size Group	Benign (n)	Malignant (n)	Total	% Malignant	χ^2 (1 df)	p	OR (95% CI)
<16 mm	13	84	97	86.6%			
≥16 mm	13	21	34	61.8%	8.26	0.004	4.0 (1.62–9.89)
Total	26	105	131	80.2%	—	—	—

Pearson χ^2 (1 df). Odds ratio compares malignancy odds for <16 mm vs. ≥16 mm.

Analysis of Bethesda category III Subtypes: Sub-analysis of Bethesda category III nodules revealed distinct sonographic profiles between subtypes. AUS-N (nuclear atypia) had higher TI-RADS scores than AUS-O (other atypia), median 4.0 (IQR 4–5) vs. 3.0 (IQR 3–4); Mann–Whitney U p = 0.002. Malignancy rates did not differ significantly on the histology subset, 87.5% (14/16) vs. 72.0% (18/25); Fisher's exact p = 0.441 (Table 4).

Table 4. Comparison of Bethesda III subtypes (AUS-N vs. AUS-O).

AUS Subtype	All Bethesda III Nodules (*n*)	Surgically Resected (Malignant/Total)	Malignancy Rate (%)	Median TI-RADS (IQR)
AUS-O (other atypia)	128	18/25	72.0	3.0 (3–4)
AUS-N (nuclear atypia)	69	14/16	87.5	4.0 (4–5)
Unspecified/not subclassified	1	—	—	—
Total	198	32/41	78.0	—

Mann–Whitney U for TI-RADS scores ($p = 0.002$); Fisher's exact (two-sided) for malignancy ($p = 0.441$). Malignancy is calculated only in the surgically resected subset.

Diagnostic Performance of Combined TI-RADS and Bethesda Thresholds: Detailed diagnostic performance analysis showed:

- The combination of TI-RADS $\geq$ 3 and Bethesda $\geq$ 5 provided perfect specificity (100%) and PPV (1.00), although with reduced sensitivity (56.2%).
- TI-RADS $\geq$ 4 and Bethesda $\geq$ 4 delivered a clinically useful rule-in balance (Sp 96.2%, PPV 98.4%, Youden J 0.552), particularly relevant to Bethesda categories III–IV triage.
- "OR" combination strategies, such as TI-RADS $\geq$ 4 or Bethesda $\geq$ 3, increased sensitivity to 95.2% but drastically reduced specificity (26.9%) and Youden index (0.222).

Overall, our analysis strongly supports a dual-modality diagnostic approach using combined TI-RADS and Bethesda thresholds, especially for Bethesda categories III–IV nodules, improving specificity and clinical decision-making precision in managing indeterminate thyroid nodules (Figure 3).

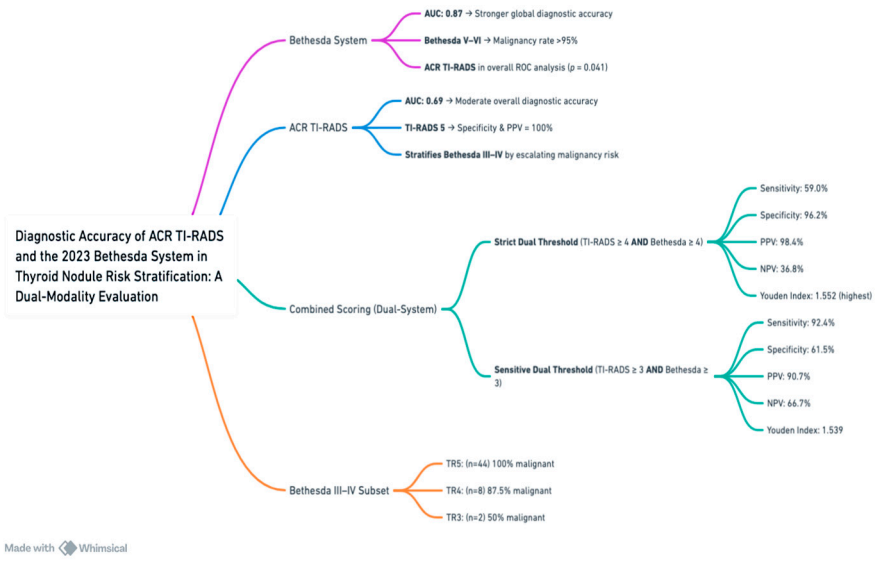

Figure 3. Conceptual framework and key diagnostic metrics for dual-modality evaluation of ACR TI-RADS and 2023 Bethesda System in thyroid nodule stratification. Conceptual framework and key diagnostic metrics for dual-modality evaluation of ACR TI-RADS and the 2023 Bethesda System. Examples: TI-RADS $\geq$ 3 AND Bethesda $\geq$ 5 (Sp 100%, PPV 100%, Se 56.2%); TI-RADS $\geq$ 4 AND Bethesda $\geq$ 4 (Sp 96.2%, PPV 98.4%, Se 59.0%); TI-RADS $\geq$ 4 OR Bethesda $\geq$ 3 (Se 95.2%, Sp 26.9%). See Supplementary Table S1 for the complete rule set.

Illustrative cases. Figure 4 presents representative multimodal examples across the spectrum from TI-RADS 3/Bethesda II (benign) to TI-RADS 5/Bethesda V (malignant), demonstrating the sonographic–cytologic–histologic progression.

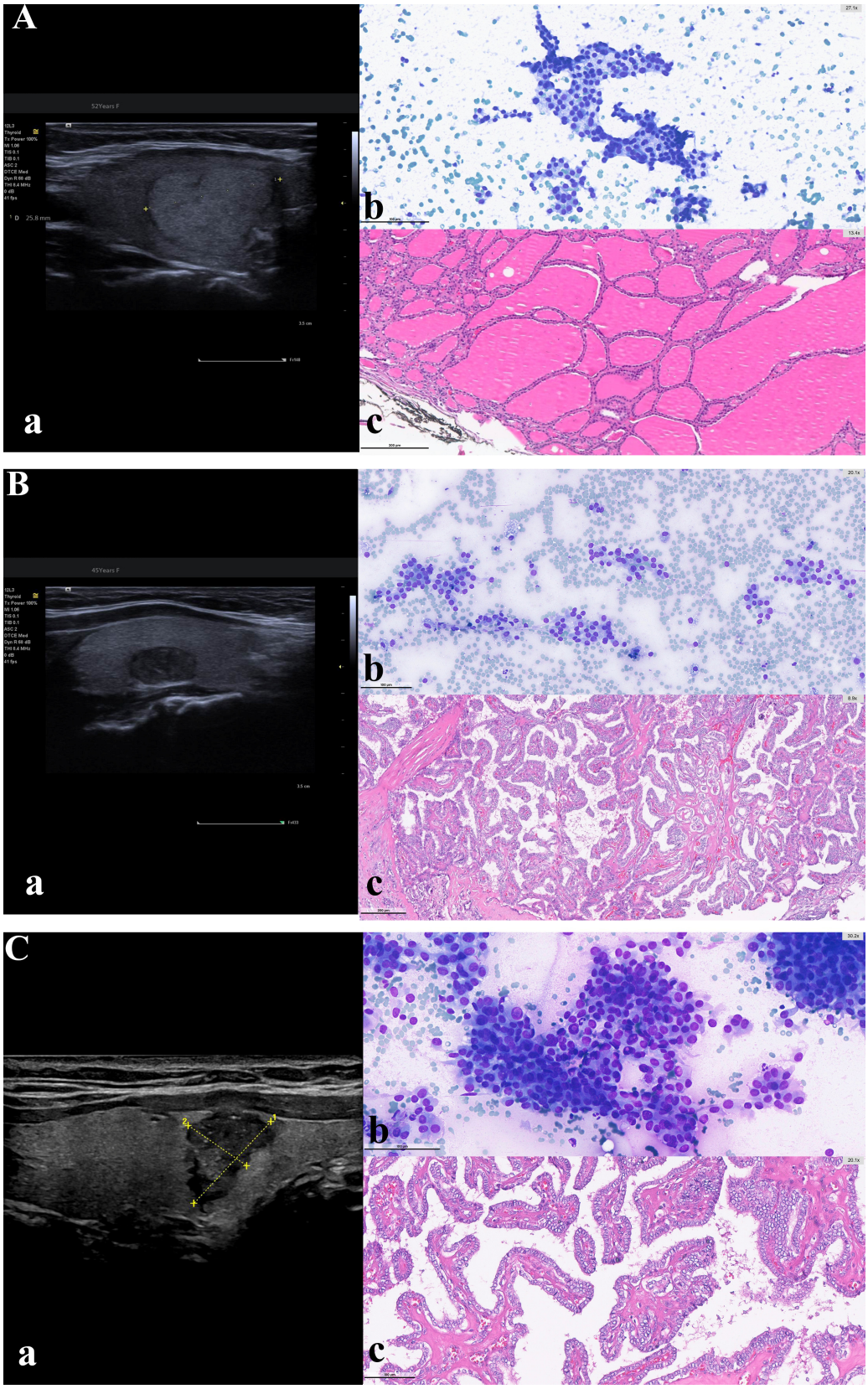

Figure 4. Radiological, cytological, and histopathological correlation of three thyroid nodules with increasing malignancy risk. (**A**) Benign nodule (TI-RADS 3/Bethesda II). (**a**) Ultrasonography shows a well-circumscribed, isoechogenic solid nodule (26 mm) surrounded by a thin hypoechoic rim in the midline of the right thyroid lobe. (**b**) Cytology reveals moderately cellular smears with monolayered

follicular cell sheets lacking atypia (MGG stain, Bethesda II). (**c**) Histopathology shows variable-sized dilated follicles lined by flattened epithelium consistent with nodular hyperplasia (H&E stain). (**B**) Indeterminate nodule with atypia (TI-RADS 4/Bethesda III). (**a**) Ultrasonography shows a 15 mm well-defined, hypoechogenic solid nodule posterior to the midline of the right lobe. (**b**) Cytology shows follicular epithelial cells with mild nuclear enlargement, irregular contours, and eccentric nuclei suggestive of atypia (MGG stain, Bethesda III—nuclear atypia). (**c**) Histopathology demonstrates complex papillary architecture and nuclear features of papillary thyroid carcinoma (H&E stain). (**C**) Malignant nodule (TI-RADS 5/Bethesda V). (**a**) Ultrasonography reveals a 15 mm irregular, hypoechogenic solid nodule in the left lobe midline, with microcalcifications and capsular disruption. (**b**) Cytology reveals papillary groups and isolated atypical epithelial cells with nuclear pseudoinclusions, grooves, and dense cytoplasm (MGG stain, Bethesda V). (**c**) Histopathology shows a classic variant of papillary thyroid carcinoma characterized by complex branching papillae, fibrovascular cores, nuclear clearing, grooves, and inclusions (H&E stain).

4. Discussion

Our study presents a prospective evaluation of the diagnostic synergy between the updated 2023 Bethesda System and the ACR TI-RADS in managing indeterminate thyroid nodules (Bethesda categories III–IV). The findings underline the added value of combining these two established frameworks to enhance risk stratification and guide clinical decision-making more precisely. The study included 1009 consecutive thyroid nodules assessed with contemporaneous ACR TI-RADS and 2023 Bethesda classification, and diagnostic accuracy analyses were based on the 131 nodules with surgical histopathologic confirmation.

In our analysis, the Bethesda System demonstrated significantly superior discriminatory performance compared to TI-RADS alone (AUC 0.87 vs. 0.69; DeLong $p = 0.041$), consistent with prior studies highlighting the higher predictive accuracy of cytological assessment over US-based classification for malignancy prediction [5,21]. In the surgically resected cohort, this stronger discriminatory contribution of cytology, together with the added value of combined TI-RADS/cytology assessment, was particularly relevant for resolving diagnostic ambiguity in Bethesda III-IV nodules and supporting more reliable interdisciplinary triage.

Within the Bethesda III–IV histology subset, malignancy risk correlated with TI-RADS category: TR3 84.2% (16/19), TR4 68.2% (15/22), and TR5 100% (7/7). These findings align with those reported by Middleton et al. [15], who emphasized that high-risk sonographic features significantly increase the likelihood of malignancy, underlining the clinical value of integrating ultrasound with cytology for more informed decision-making. Historically, ROM for AUS/Bethesda III has been reported as <15%, although several large cohorts have documented rates approaching 50% [6,7,9–11,22,23]. Indeterminate categories (Bethesda III–IV) account for roughly one quarter of FNAs [6,7], and a uniform management standard is still lacking. Our substantially higher ROM likely reflects the surgically enriched cohort and tightly standardized interpretations in our setting. Similar verification-bias effects have been reported in tertiary-referral AUS/FLUS series, where excision-based malignancy rates are much higher than enrollment-based rates [23]. However, the updated 2023 Bethesda criteria (including refined definitions of atypia and AUS subtyping) may also shift the case mix classified as Bethesda III compared with older series, potentially contributing to higher observed ROM [7]. Our design does not allow us to disentangle the effect of classification updates from surgical selection/referral patterns in explaining the high Bethesda III ROM.

A dual-threshold strategy performed best for rule-in. The combination TI-RADS ≥ 4 AND Bethesda ≥ 4 achieved Se 59.0%, Sp 96.2%, PPV 98.4%, and a high Youden index (J = 0.552), reflecting prior observations that adding targeted US features to cytology boosts specificity and PPV with minimal sensitivity loss [22,24,25]. By contrast, OR strategies traded specificity for sensitivity; for example, TI-RADS ≥ 4 OR Bethesda ≥ 3 yielded

Se 95.2% but Sp 26.9% (J = 0.222), and TI-RADS $\geq$ 4 OR Bethesda $\geq$ 4 yielded Se 81.0%, Sp 38.5%. Although some OR rules (e.g., 3 OR 3) showed very high NPV in this surgically enriched subset, their very low specificity limits rule-out utility and generalizability. Overall, the selected AND rules function best as confirmatory rule-in thresholds (see Figure 2 and Supplementary Table S1).

These metrics highlight the optimal balance between sensitivity and specificity, supporting prior literature that advocates combining US risk stratification systems with cytological interpretation to enhance diagnostic precision and reduce unnecessary surgeries [13,14]. Most comparative TI-RADS studies report performance at single risk-category cut-offs (e.g., 'biopsy all TI-RADS $\geq$ 4') without incorporating the diameter thresholds that guideline algorithms mandate (e.g., TI-RADS 4 nodules $\geq$ 15 mm; TI-RADS 5 nodules $\geq$ 10 mm). By contrast, our analysis applies these size criteria alongside category scores, thereby more closely indicating the recommended clinical workflow and offering a more realistic assessment of diagnostic accuracy [4,12,24]. This context supports our choice to analyze both category cutoffs and combined criteria. In a recent Bethesda III-IV cohort, ACR TI-RADS showed the highest specificity and NPV, and achieved the highest AUC (0.842) and PPV, underlining that US systems trade sensitivity for specificity differently [4,24–27]. Our specificity of 96.2% and PPV of 98.4% at TI-RADS $\geq$ 4/Bethesda $\geq$ 4 mirrors that high-specificity profile.

Comparable high proportions have been noted elsewhere: one series reported 71% malignancy in TR4 and 100% in TR5, but emphasized the very small surgical subset (n = 10) [22] in malignancy from 0% in TR1 to 88.8% in TR5; although no TR1 nodules were present in our cohort, we similarly observed progressively higher malignancy rates with increasing TI-RADS categories [22,28].

In the context of TI-RADS and Bethesda classification, nodule diameter provided only marginal additional predictive value. In the full histology cohort, smaller nodules (<16 mm) were more often malignant than larger ones (86.6% vs. 61.8%; p = 0.004; OR 4.0; Table 3). This inverse size–malignancy association should be interpreted cautiously, as selective biopsy and surgical referral of sonographically and/or cytologically suspicious small nodules may enrich the smaller-size group for malignancy. Although a component-level sonographic feature breakdown was not available, ACR TI-RADS categories are derived from these features, and the surgically confirmed cohort was predominantly TI-RADS TR4–TR5 (97/131, 74%), consistent with preferential work-up of sonographically suspicious nodules, including smaller nodules. Size alone performed poorly as a classifier (AUROC 0.66). When we narrowed to Bethesda III–IV, the effect disappeared (77.1% vs. 78.6%; p = 1.00; OR 0.92; see Supplementary Table S3), underlining how unstable size is once cytologic and sonographic morphology render a nodule "indeterminate". Others have reported similar findings: a recent series of Bethesda III nodules found no size–ROM association (p = 0.55) [25], and a meta-analysis of indeterminate lesions showed higher malignancy in nodules <4 cm (37% vs. 15%), albeit with extreme between-study heterogeneity (I^2 = 99%) [16]. In that context, the high size thresholds in ACR TI-RADS make sense, as they were designed to reduce benign FNAs, not to declare small nodules biologically safe. This approach is consistent with findings from large real-world registries: ACR TI-RADS recommends biopsy in only ~29% of 27,933 nodules, 25–50% fewer than ATA, K TI-RADS, or EU TI-RADS [18]. The flip side is that a few cancers will fall below those size triggers: authors warn that the TR3 2.5 cm cut-off can miss small tumors, and Huang et al. found 5.7% of TR4 and 25% of TR5 cancers were <1.5 cm [17]. Sonographic and cytologic morphology should remain the primary drivers of escalation; size should be interpreted as a procedural threshold variable, not an independent risk marker. By prioritizing sonographic and cytologic morphology over size thresholds alone, this approach may facilitate earlier identification and management of clinically significant nodules that would otherwise fall below biopsy cut-offs, potentially reducing delayed presentation with

very large ('giant') thyroid tumors described in surgical series; however, this outcome was not assessed in the present study [29].

Despite significantly higher median TI-RADS scores in AUS-N compared with AUS-O (median 4.0 [IQR 4–5] vs. 3.0 [IQR 3–4]; Mann–Whitney $p = 0.002$), malignancy rates did not differ (87.5% [14/16] vs. 72.0% [18/25]; Fisher's $p = 0.441$). See Table 4 and Supplementary Table S4. This pattern reflects older reports using the FLUS/architectural atypia category (now equivalent to AUS-O in the 2023 Bethesda System): for example, Baser et al. documented an AUC of 0.58 for TI-RADS $\geq$ 4c in AUS-N with no meaningful discrimination in FLUS [11], and Słowińska Klencka et al. found EU TI-RADS outperformed repeat FNA in AUS-N but not in architectural-atypia/FLUS [10]. Moreover, standalone TI-RADS performance in our cohort was modest in AUS-N (AUC 0.61) and poor in AUS-O (AUC 0.44). Therefore, cytologic subtype remains an essential context, and TI-RADS should augment, but not replace, cytological assessment.

Finally, applying a dual threshold of TI-RADS $\geq$ 4 together with Bethesda $\geq$ 4 identified a high-risk subset in the full cohort (specificity 96.2%, PPV 98.4%, Youden index 0.552; see Figure 2 and Supplementary Table S1).

The histologically confirmed subset was enriched for malignancy due to surgical selection bias, as surgery was preferentially performed in nodules with higher clinical/sonographic/cytologic suspicion, which may have inflated the positive predictive value (PPV) and overall ROM; sensitivity/specificity are less affected by prevalence. Accordingly, PPV is likely overestimated and NPV underestimated in this surgically enriched cohort relative to an unselected FNA population, limiting the generalizability of prevalence-dependent predictive values. Ultrasound data were captured primarily as overall ACR TI-RADS categories rather than individual component features (composition, echogenicity, shape, margins, and echogenic foci), which limited feature-level comparisons between size groups. The consistently high ROM observed, particularly in indeterminate categories, likely reflects standardized diagnostic workflows implemented by an experienced radiologist and cytopathologist, a methodological strength that can also introduce interpretive bias. Because borderline entities such as NIFTP, WDT-UMP, and FT-UMP commonly arise from indeterminate cytology, their exclusion may increase the apparent ROM within Bethesda III–IV compared with series that include these outcomes in the non-malignant/borderline denominator. Published series comparing TI-RADS frameworks are largely retrospective with small surgically confirmed subsets and notable interobserver variability, factors that likely explain the wide ROM ranges (e.g., 29–82% across TR4/4c) [22,25,27]. Our prospective, single-team design reduces some variability but simultaneously enriches ROM; this consistently high ROM, especially within indeterminate categories, likely reflects expert, standardized protocols (single radiologist/cytopathologist) and a surgically enriched cohort, phenomena also noted in comparative TI-RADS studies [6,22,25,27]. Molecular testing, which could further stratify risk, was not included due to logistical constraints and represents a direction for future refinement. In current clinical practice, molecular testing is an important adjunct for Bethesda III–IV nodules, particularly for rule-out strategies and to reduce potentially avoidable surgery [6,28,30]. Our combined TI-RADS/Bethesda approach is not proposed as a replacement for molecular testing; rather, it may complement molecular testing by identifying higher-risk nodules for escalation and may provide pragmatic risk stratification when molecular testing is unavailable or resource-limited. Future studies should evaluate combined imaging–cytology strategies alongside molecular testing in broader, less surgically enriched cohorts. While the single-center design may limit generalizability, this is counterbalanced by the prospective nature of data collection, the use of uniform reporting criteria (2023 Bethesda System and ACR TI-RADS), and the

consistency of interpretation across imaging and cytology, which collectively strengthen internal validity.

In light of its prospective design, standardized diagnostic protocols, and integrated analytical approach, this study provides clinically relevant evidence supporting the use of combined cytological and sonographic criteria to improve risk stratification and guide management decisions in indeterminate thyroid nodules.

5. Conclusions

This prospective study strengthened the diagnostic superiority of combining ACR TI-RADS and Bethesda scoring for Bethesda III–IV thyroid nodules. The dual-strategy approach significantly enhances diagnostic specificity and risk stratification, addressing critical limitations when relying solely on cytological or ultrasonographic assessments. Our findings advocate a morphology-driven approach over traditional size-based decision-making, emphasizing the importance of sonographic features in optimizing clinical management strategies. Future guidelines and clinical protocols should consider incorporating combined diagnostic thresholds as standard practice to reduce unnecessary interventions and improve patient outcomes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/endocrines7010012/s1, Table S1: Top dual-threshold rules by Youden (histology subset; Table S2: AUROC Comparison of Predictors for Malignancy; Table S3: 16 mm cut-off applied only to Bethesda III–IV nodules (n = 49); Table S4: Comparison of Bethesda category III Subtypes (AUS-N vs. AUS-O).

Author Contributions: Conceptualization, O.A. and S.C.; methodology, S.C., O.A., M.G.G. and B.C.; software, S.C., M.S., J.P.S. and M.G.G.; validation, O.A., B.C. and S.C.; formal analysis, S.C. and M.G.G.; investigation, O.A., B.C., C.K.K., M.S. and S.C.; resources, O.A., B.C. and C.K.K.; data curation, O.A., B.C. and C.K.K.; writing—original draft preparation, O.A. and S.C.; writing—review and editing, O.A., B.C., C.K.K., M.G.G., M.S., J.P.S. and S.C.; visualization, O.A., B.C., C.K.K., M.S., J.P.S. and S.C.; supervision, S.C.; project administration, O.A. and S.C.; funding acquisition, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Acibadem University (protocol code 2025-06/39 and date of approval: 17 April 2025).

Informed Consent Statement: This study is retrospective in design; no biological samples were collected, and no additional analyses were performed. Only anonymized data obtained from the Department of Pathology medical records were used. All personal, medical, and social information of the patients was strictly kept confidential and was not shared. Therefore, no patient informed consent documentation is required for this study.

Data Availability Statement: The datasets generated and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vaccarella, S.; Franceschi, S.; Bray, F.; Wild, C.P.; Plummer, M.; Dal Maso, L. Worldwide Thyroid-Cancer Epidemic? The Increasing Impact of Overdiagnosis. *N. Engl. J. Med.* **2016**, *375*, 614–617. [CrossRef] [PubMed]
2. Esserman, L.J.; Thompson, I.M., Jr.; Reid, B. Overdiagnosis and overtreatment in cancer: An opportunity for improvement. *JAMA* **2013**, *310*, 797–798. [CrossRef] [PubMed]
3. Hoang, J.K.; Nguyen, X.V. Understanding the Risks and Harms of Management of Incidental Thyroid Nodules: A Review. *JAMA Otolaryngol. Head Neck Surg.* **2017**, *143*, 718–724. [CrossRef] [PubMed]

4. Tessler, F.N.; Middleton, W.D.; Grant, E.G.; Hoang, J.K.; Berland, L.L.; Teefey, S.A.; Cronan, J.J.; Beland, M.D.; Desser, T.S.; Frates, M.C.; et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J. Am. Coll. Radiol.* **2017**, *14*, 587–595. [CrossRef] [PubMed]
5. Li, W.; Wang, Y.; Wen, J.; Zhang, L.; Sun, Y. Diagnostic Performance of American College of Radiology TI-RADS: A Systematic Review and Meta-Analysis. *AJR Am. J. Roentgenol.* **2021**, *216*, 38–47. [CrossRef] [PubMed]
6. Ali, S.Z.; Baloch, Z.W.; Cochand-Priollet, B.; Schmitt, F.C.; Vielh, P.; VanderLaan, P.A. The 2023 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid* **2023**, *33*, 1039–1044. [CrossRef] [PubMed]
7. Guerreiro, S.C.; Tastekin, E.; Mourao, M.; Loureiro, I.; Eusebio, R.; Marques, H.P.; Oznur, M.; Caliskan, C.K.; Schmitt, F.C.; Bongiovanni, M.; et al. Impact of the 3rd Edition of the Bethesda System for Reporting Thyroid Cytopathology on Grey Zone Categories. *Acta Cytol.* **2023**, *67*, 593–603. [CrossRef] [PubMed]
8. Turkdogan, S.; Pusztaszeri, M.; Forest, V.I.; Hier, M.P.; Payne, R.J. Are Bethesda III Thyroid Nodules More Aggressive than Bethesda IV Thyroid Nodules When Found to Be Malignant? *Cancers* **2020**, *12*, 2563. [CrossRef] [PubMed] [PubMed Central]
9. Belovarac, B.; Zhou, F.; Modi, L.; Sun, W.; Shafizadeh, N.; Negron, R.; Yee-Chang, M.; Szeto, O.; Simsir, A.; Sheth, S.; et al. Evaluation of ACR TI-RADS cytologically indeterminate thyroid nodules and molecular profiles: A single-institutional experience. *J. Am. Soc. Cytopathol.* **2022**, *11*, 165–172. [CrossRef] [PubMed]
10. Słowińska-Klencka, D.; Popowicz, B.; Duda-Szymańska, J.; Klencki, M. Thyroid Nodules with Nuclear Atypia of Undetermined Significance (AUS-Nuclear) Hold a Two-Times-Higher Risk of Malignancy than AUS-Other Nodules Regardless of EU-TIRADS Class of the Nodule or Borderline Tumor Interpretation. *Cancers* **2025**, *17*, 1365. [CrossRef] [PubMed] [PubMed Central]
11. Baser, H.; Cakir, B.; Topaloglu, O.; Alkan, A.; Polat, S.B.; Dogan, H.T.; Yazicioğlu, M.O.; Aydin, C.; Ersoy, R. Diagnostic accuracy of Thyroid Imaging Reporting and Data System in the prediction of malignancy in nodules with atypia and follicular lesion of undetermined significance cytologies. *Clin. Endocrinol.* **2017**, *86*, 584–590. [CrossRef] [PubMed]
12. Kim, D.H.; Kim, S.W.; Basurrah, M.A.; Lee, J.; Hwang, S.H. Diagnostic Performance of Six Ultrasound Risk Stratification Systems for Thyroid Nodules: A Systematic Review and Network Meta-Analysis. *AJR Am. J. Roentgenol.* **2023**, *220*, 791–803. [CrossRef] [PubMed]
13. Eissa, M.S.; Sabry, R.M.; Abdellateif, M.S. Evaluating the Diagnostic Role of ACR-TIRADS and Bethesda Classifications in Thyroid Nodules Highlighted by Cyto-Histopathological Studies. *Exp. Clin. Endocrinol. Diabetes* **2024**, *132*, 596–606. [CrossRef] [PubMed]
14. Ramonell, K.M.; Ohori, N.P.; Liu, J.B.; McCoy, K.L.; Furlan, A.; Tublin, M.; Carty, S.E.; Yip, L. Changes in thyroid nodule cytology rates after institutional implementation of the Thyroid Imaging Reporting and Data System. *Surgery* **2023**, *173*, 232–238. [CrossRef] [PubMed] [PubMed Central]
15. Middleton, W.D.; Teefey, S.A.; Tessler, F.N.; Hoang, J.K.; Reading, C.C.; Langer, J.E.; Beland, M.D.; Szabunio, M.M.; Desser, T.S. Analysis of Malignant Thyroid Nodules That Do Not Meet ACR TI-RADS Criteria for Fine-Needle Aspiration. *AJR Am. J. Roentgenol.* **2021**, *216*, 471–478. [CrossRef] [PubMed]
16. Cotter, A.; Jinih, M. Thyroid nodule size and risk of malignancy: A systematic review. *Discov. Oncol.* **2025**, *16*, 1188. [CrossRef] [PubMed] [PubMed Central]
17. Huang, E.Y.F.; Kao, N.H.; Lin, S.Y.; Jang, I.J.H.; Kiong, K.L.; See, A.; Venkatanarasimha, N.; Lee, K.A.; Lim, C.M. Concordance of the ACR TI-RADS Classification with Bethesda Scoring and Histopathology Risk Stratification of Thyroid Nodules. *JAMA Netw. Open* **2023**, *6*, e2331612. [CrossRef] [PubMed] [PubMed Central]
18. Alqahtani, S.M.; Albalawi, H.I.; Alalawi, Y.S.; AlFattani, A.A.; Al-Sobhi, S.S. The impact of nodule size on malignancy risk in indeterminate thyroid nodules. *Gland Surg.* **2024**, *13*, 470–479. [CrossRef] [PubMed] [PubMed Central]
19. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA* **2013**, *310*, 2191–2194. [CrossRef]
20. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). 2016. Available online: https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf (accessed on 28 February 2026).
21. Modi, L.; Sun, W.; Shafizadeh, N.; Negron, R.; Yee-Chang, M.; Zhou, F.; Simsir, A.; Sheth, S.; Brandler, T.C. Does a higher American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) score forecast an increased risk of malignancy? A correlation study of ACR TI-RADS with FNA cytology in the evaluation of thyroid nodules. *Cancer Cytopathol.* **2020**, *128*, 470–481. [CrossRef] [PubMed]
22. Guo, J.; Du, L.; Bi, W.; Liu, Y.; Zhang, C. Retrospective study from a single center to comparison of diagnostic value of three thyroid imaging reporting and data systems in Bethesda III/IV thyroid nodules. *Front. Oncol.* **2025**, *15*, 1549646. [CrossRef] [PubMed] [PubMed Central]
23. Kim, T.H.; Jeong, D.J.; Hahn, S.Y.; Shin, J.H.; Oh, Y.L.; Ki, C.S.; Kim, J.W.; Jang, J.Y.; Cho, Y.Y.; Chung, J.H.; et al. Triage of patients with AUS/FLUS on thyroid cytopathology: Effectiveness of the multimodal diagnostic techniques. *Cancer Med.* **2016**, *5*, 769–777. [CrossRef]

24. Yang, L.; Li, C.; Chen, Z.; He, S.; Wang, Z.; Liu, J. Diagnostic efficiency among Eu-/C-/ACR-TIRADS and S-Detect for thyroid nodules: A systematic review and network meta-analysis. *Front. Endocrinol.* **2023**, *14*, 1227339. [CrossRef] [PubMed] [PubMed Central]
25. Seminati, D.; Capitoli, G.; Leni, D.; Fior, D.; Vacirca, F.; Di Bella, C.; Galimberti, S.; L'Imperio, V.; Pagni, F. Use of Diagnostic Criteria from ACR and EU-TIRADS Systems to Improve the Performance of Cytology in Thyroid Nodule Triage. *Cancers* **2021**, *13*, 5439. [CrossRef] [PubMed] [PubMed Central]
26. Sarayu, S.; Nair, A.; Khader, J.P.; Rema, S.P.P.; Meerasainaba, S.; Kumar, S.; Gomez, R.; Chellamma, J. Prospective Validation of Accuracy of American College of Radiologists- Thyroid Imaging Reporting and Data System (ACR-TIRADS) in Diagnosing Malignancy in Thyroid Nodule and a Prediction Score (TiPS) for Thyroid Malignancy. *Indian J. Endocrinol. Metab.* **2025**, *29*, 101–107. [CrossRef] [PubMed] [PubMed Central]
27. Castilla Villanueva, M.Á.; Solis Cano, D.G.; Amador Martínez, A.; Téliz Meneses, M.A.; Baquera-Heredia, J.; Vallin Orozco, C.E.; Loya Ceballos, M. Individual Ultrasonographic Characteristics of Thyroid Nodules and Their Cytopathological Correlation to Determine Malignancy Risk. *Cureus* **2024**, *16*, e63918. [CrossRef] [PubMed] [PubMed Central]
28. Tirrò, E.; Martorana, F.; Romano, C.; Vitale, S.R.; Motta, G.; Di Gregorio, S.; Massimino, M.; Pennisi, M.S.; Stella, S.; Puma, A.; et al. Molecular Alterations in Thyroid Cancer: From Bench to Clinical Practice. *Genes* **2019**, *10*, 709. [CrossRef] [PubMed] [PubMed Central]
29. Vrinceanu, D.; Dumitru, M.; Marinescu, A.; Serboiu, C.; Musat, G.; Radulescu, M.; Popa-Cherecheanu, M.; Ciornei, C.; Manole, F. Management of Giant Thyroid Tumors in Patients with Multiple Comorbidities in a Tertiary Head and Neck Surgery Center. *Biomedicines* **2024**, *12*, 2204. [CrossRef] [PubMed]
30. Durante, C.; Hegedüs, L.; Czarniecka, A.; Paschke, R.; Russ, G.; Schmitt, F.; Soares, P.; Solymosi, T.; Papini, E. 2023 European Thyroid Association Clinical Practice Guidelines for thyroid nodule management. *Eur. Thyroid J.* **2023**, *12*, e230067.