International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	ARTIFICIAL INTELLIGENCE FOR ULTRASOUND-BASED DIAGNOSIS AND RISK STRATIFICATION OF THYROID NODULES: EVIDENCE, HUMAN FACTORS, AND HEALTH-IT IMPLICATIONS
DOI	https://doi.org/10.31435/ijitss.2(50).2026.5276
RECEIVED	03 February 2026
ACCEPTED	05 May 2026
PUBLISHED	14 May 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

ARTIFICIAL INTELLIGENCE FOR ULTRASOUND-BASED DIAGNOSIS AND RISK STRATIFICATION OF THYROID NODULES: EVIDENCE, HUMAN FACTORS, AND HEALTH-IT IMPLICATIONS

Joanna Ślusarczyk *(Corresponding Author, Email: jlusarczyk11@gmail.com)*
Provincial Hospital in Kielce, Poland
ORCID ID: 0009-0008-6405-5963

Beata Szreder
Medical University of Gdansk, Gdansk, Poland
ORCID ID: 0009-0008-7876-2159

Piotr Pietrzyk
Mazovian Bródno Hospital in Warsaw, Poland

Natalia Stanek
Hospital of the Ministry of Interior and Administration in Katowice, Poland
ORCID ID: 0009-0009-5745-5253

Patrycja Krawczyk
Provincial Hospital in Kielce, Poland
ORCID ID: 0009-0001-0536-4570

Maja Łapaj
Hospital of the Ministry of Interior and Administration in Katowice, Poland
ORCID ID: 0009-0006-8285-6997

Aleksandra Jagiełło
Mazovian Bródno Hospital in Warsaw, Poland
ORCID ID: 0009-0005-6310-6755

Zuzanna Noweta
Collegium Medicum of the Nicolaus Copernicus University in Bydgoszcz, Poland
ORCID ID: 0009-0009-8086-8101

Milena Lewicka
Medical University of Gdansk, Gdansk, Poland
ORCID ID: 0009-0006-2919-9537

Tamara Chodań
Collegium Medicum of the Nicolaus Copernicus University in Bydgoszcz, Poland
ORCID ID: 0009-0001-5891-7925

ABSTRACT

Background: Thyroid ultrasound is the cornerstone of thyroid nodule assessment, yet image capture and interpretation remain operator-dependent. Artificial intelligence (AI) is increasingly proposed to support malignancy risk stratification, align reporting with TI-RADS frameworks, and reduce unnecessary fine-needle aspiration biopsy (FNA).
Objective: To synthesize evidence from the last five years on AI-assisted thyroid nodule assessment using ultrasound, with emphasis on validation, generalizability, human–AI interaction, and health-IT implications.
Methodology: Structured narrative review prioritizing multicenter or externally validated studies, workflow-oriented evaluations, and systematic reviews/meta-analyses.
Results: AI may improve consistency among less experienced users in retrospective studies, but a prospective trial showed no workflow improvement and strong operator dependence. Recent multicenter studies increasingly use multi-view images, cine/video, and multimodal copilots. Meta-analyses highlight heterogeneity and the need for robust external validation.
Conclusion: Translation requires prospective decision-endpoint studies, standardized ultrasound scanning, uncertainty-aware outputs, and integration into structured reporting and monitoring systems.

KEYWORDS

Thyroid Nodules, Ultrasound, Artificial Intelligence, Deep Learning, Computer-Aided Diagnosis, TI-RADS

CITATION

Joanna Ślusarczyk, Beata Szreder, Piotr Pietrzyk, Natalia Stanek, Patrycja Krawczyk, Maja Łapaj, Aleksandra Jagiełło, Zuzanna Noweta, Milena Lewicka, Tamara Chodań. (2026) Artificial Intelligence for Ultrasound-Based Diagnosis and Risk Stratification of Thyroid Nodules: Evidence, Human Factors, and Health-IT Implications. *International Journal of Innovative Technologies in Social Science*. 2(50). doi: 10.31435/ijitss.2(50).2026.5276

1. Introduction

Thyroid nodules are among the most common findings in endocrine and head-and-neck practice. High-resolution ultrasound (US) now detects nodules readily, and detection is often incidental. The clinical challenge is triage (decision-making about biopsy vs follow-up): which nodules warrant biopsy, which can be safely observed, and which are so low risk that continued follow-up is more harmful than helpful.

Standardized risk stratification systems, including ACR TI-RADS and EU-TIRADS, provide structured descriptors and size thresholds to link ultrasound appearance to management (Tessler et al., 2017; Russ et al., 2017). ATA guideline patterns are also widely used in endocrine pathways (Haugen et al., 2016). These frameworks improve communication, but uncertainty persists because many sonographic features are subjective and ultrasound scanning is operator-dependent.

Operator dependence is not a minor limitation—it is fundamental. Probe pressure, angle, gain, and plane selection influence appearance and feature visibility. Therefore, interobserver variability is partly a capture problem, not only a reading problem. This matters for AI because the model's input is the operator's chosen representation of the nodule.

AI for thyroid ultrasound has evolved from classic computer-aided diagnosis (CAD) to deep learning models and, more recently, multimodal generative "copilots" (Yao et al., 2025). This shift has coincided with growing regulatory interest and commercial availability (Wildman-Tobriner et al., 2022; See-Mode Technologies, 2024).

For IJITSS audiences, thyroid ultrasound AI is a useful sociotechnical (involving people, workflow, and technology together) change: a digital tool embedded in workflow can change clinician behavior, biopsy rates, and patient experience. The question is not only whether AI is accurate, but whether it improves decisions safely, equitably, and in a way that can be audited and governed over time.

This review brings together evidence from the last five years and highlights implications for clinical pathways, health information systems, governance (formal oversight, policies, and accountability), and future research design.

2. Methodology

2.1 Review design. This article is a structured narrative review of artificial intelligence (AI) for ultrasound-based assessment and risk stratification of thyroid nodules. The goal is to synthesize clinically relevant evidence—especially validation, generalizability, and human–AI interaction—rather than to list every model variation.

2.2 Evidence identification. We performed targeted searches and citation chaining in PubMed and on publisher websites, prioritizing peer-reviewed studies published over the last five years. Search terms covered thyroid nodules, ultrasound, artificial intelligence/deep learning, computer-aided diagnosis, TI-RADS, and named systems (e.g., S-Detect). Key guideline documents (ATA, ACR TI-RADS, EU-TIRADS) were used to frame clinical context.

2.3 Inclusion emphasis. We prioritized multicenter and/or externally validated studies, prospective or workflow-oriented evaluations, dataset papers that support reproducibility, and systematic reviews/meta-analyses. Studies were favored when they clearly defined reference standards (cytology/histopathology and/or follow-up), minimized patient-level leakage in data splits, and reported performance beyond internal testing.

2.4 Interpretation framework. We interpreted findings using established reporting and appraisal approaches commonly used in diagnostic AI research, focusing on recurring translational limitations such as spectrum bias, data leakage, domain shift, and incomplete reporting.

3. Results

3.1 Clinical context: where AI fits in TI-RADS pathways

AI can support TI-RADS workflows at multiple points: feature extraction, category suggestion, biopsy triage support, and structured reporting. The most meaningful way to evaluate AI is not only AUC but the impact on decisions: how AI changes biopsy recommendations and biopsy yield (malignancies per biopsies performed) under defined thresholds.

3.2 Commercial CAD and workflow evidence

Retrospective studies suggest CAD may help less experienced users and improve agreement (Kang et al., 2021; Huang et al., 2022). In contrast, a prospective clinical trial found that CAD did not improve diagnostic evaluation (evaluation) among novice and intermediate operators and that AI accuracy depended strongly on the operator (Edström et al., 2025).

Regulatory summaries such as SMART-T emphasize AI as decision support and caution against basing management solely on AI outputs (See-Mode Technologies, 2024).

3.3 Deep learning, cine/video, and multimodal copilots

Multicenter deep learning approaches increasingly use multi-view inputs and clinically aligned intermediate tasks (Vahdati et al., 2024; Chen et al., 2024). Cine/video systems such as TNVis aim to reflect dynamic scanning and reduce dependence on frame selection (Zhou et al., 2025). Multimodal copilots such as ThyGPT add report-aware and guideline-aware assistance but introduce new governance risks related to generated text (Yao et al., 2025).

3.4 Evidence syntheses

Meta-analyses consistently report promising diagnostic performance but also stress heterogeneity and the importance of external validation (Gatta et al., 2025; Zhan et al., 2025; Xu et al., 2023; Potipimpanon et al., 2022; Ni et al., 2025).

Finally, a head-to-head meta-analysis comparing AI with radiologists in ultrasound diagnosis found broadly comparable AUC between AI systems and radiologists, while also suggesting that deep learning approaches tend to outperform classic machine-learning approaches in subgroup analyses (Potipimpanon et al., 2022).

Two additional meta-analyses help interpret how deep learning performs when external validation is required. Xu et al. (2023) restricted inclusion to epidemiological studies that reported independent external test sets and showed that pooled performance is typically lower on external validation than internal testing, but remains clinically promising overall.

Recent evidence syntheses also add more granular insight into what AI is good at in thyroid ultrasound—beyond benign/malignant classification. Ni et al. (2025) conducted a systematic review and meta-analysis focused specifically on deep learning for thyroid nodule detection and segmentation tasks, reporting pooled performance for both task families and highlighting heterogeneity across model designs and datasets.

4. Discussion

4.1 Workflow-oriented evidence can contradict retrospective accuracy because retrospective datasets often contain selected high-quality frames and enriched malignant prevalence. Prospective evaluations, such as Edström et al. (2025), demonstrate that live scanning introduces ultrasound scanning variability that can dominate AI behavior.

4.2 Operator dependence changes the implementation problem: ultrasound scanning protocols, training, and quality control become part of the intervention. Multi-view or cine/video inputs can reduce but not eliminate sensitivity to frame selection.

4.3 Choosing the role of AI (second reader, reporting assistant, triage, copilot) determines risk. A staged adoption strategy—starting with second-reader/reporting support—can reduce harm while building monitoring capability.

4.4 Thresholds and calibration (how well predicted risks match observed outcomes) determine real-world outcomes. Institutions should validate calibration locally and choose thresholds aligned with follow-up reliability and risk tolerance.

4.5 Structured, versioned outputs linked to images enable auditing and drift (changes in model behavior/performance over time) monitoring; without this health-IT layer, accountability and safety monitoring are weak.

4.6 Ultrasound scanning quality assurance and "AI-ready" scanning protocols. One practical lesson from the live-scanning trial by Edström et al. (2025) is that variability in what the AI receives can dominate the outcome. For adoption, institutions should define a minimal ultrasound scanning protocol that makes scans "AI-ready." In thyroid nodule workflows, this may include standardized transverse and longitudinal views, documentation of the nodule in orthogonal planes, and—where feasible—short cine loops that capture the nodule margins and echogenic foci across angles. Quality assurance can be implemented as lightweight checklists embedded in the reporting process. Importantly, QA is not about making ultrasound rigid. it is about ensuring that the images used for decision support are sufficiently representative. A useful mental model is that an AI tool is a measurement instrument: it requires calibration (threshold setting) and it requires a stable measurement process (ultrasound scanning standards). In practice, this can be operationalized by periodic peer review of scans used for AI inference, feedback sessions for operators, and defining a process for handling low-quality images (e.g., rescanning, second opinion, or "AI not applicable" labeling). This approach also improves human performance, because good ultrasound scanning standards help clinicians see the same features more consistently, independent of AI.

4.7 Training and competency: using AI without letting it replace learning. When AI is introduced into a workflow, there is a risk that trainees will "delegate thinking" to the tool. In thyroid ultrasound, this risk is amplified because TI-RADS descriptors require deliberate practice and because subtle feature recognition improves with experience. A safer educational model frames AI as a feedback instrument rather than an oracle. For example, trainees can perform an initial assessment, then compare with AI-suggested descriptors and review disagreements with a supervisor. Over time, this can improve consistency and reduce overconfidence. Retrospective evidence that CAD narrows the gap between novice and expert performance (Kang et al., 2021; Huang et al., 2022) is compatible with this teaching use-case. However, the Edström et al. (2025) findings suggest that novices may not benefit automatically in live scanning. therefore, training should include ultrasound scanning coaching and explicit instruction on how to interpret AI confidence. Competency frameworks could include (a) ability to acquire TI-RADS-relevant views, (b) ability to recognize when AI is likely unreliable (e.g., poor visualization, heavy artifact), and (c) ability to justify decisions in guideline language even when AI suggests a different category.

4.8 Calibration workshops and threshold governance in real institutions. Even when a model is accurate on average, institutions must decide how to use its output. This is essentially a governance problem: setting thresholds for action, defining when AI is advisory versus when it triggers mandatory review, and aligning AI with the institution's risk tolerance. One practical approach is a "calibration workshop" during pilot deployment. In such a workshop, clinicians review a set of recent consecutive cases with known outcomes, compare human and AI outputs, and discuss threshold scenarios: what would change if the institution followed AI when it disagreed with the clinician? How many biopsies would be avoided? How many malignant nodules would be delayed? These discussions convert abstract AUC numbers into concrete pathway implications. They also reveal differences in risk tolerance across specialties (radiology, endocrinology, surgery). Because TI-RADS already encodes thresholds, a pragmatic option is to restrict AI to suggesting TI-RADS descriptors and categories, leaving size-threshold decisions to clinicians. Over time, if monitoring shows stable benefit,

institutions can adjust the role of AI. This staged governance approach is more realistic than assuming a single "optimal" threshold exists for all settings.

4.9 Handling indeterminate cytology and mixed reference standards. In thyroid nodule care, indeterminate cytology is a major driver of repeat procedures and patient stress. Yet many AI studies either exclude indeterminate nodules or label them inconsistently because definitive histopathology is available only for a subset. For translation, two practical steps are helpful. First, institutions should evaluate AI performance separately in cases that correspond to indeterminate pathways (e.g., nodules with equivocal ultrasound features or Bethesda III/IV categories when available). Second, monitoring should track not only malignancy outcomes but also how AI affects downstream choices: repeat FNA, molecular testing, or surgery. A tool that shifts more nodules into "biopsy" categories could increase indeterminate cases and thus increase downstream interventions, even if its malignancy classification accuracy is respectable. Conversely, a tool that reduces biopsies might shift uncertainty to follow-up imaging. These are pathway-level outcomes and should be part of evaluation. Meta-analyses that focus on pooled sensitivity and specificity (Gatta et al., 2025; Zhan et al., 2025) are useful, but they rarely capture the complexity of indeterminate pathways. Prospective studies designed around these pathways are needed to determine whether AI reduces or merely redistributes uncertainty.

4.10 Procurement and vendor evaluation: moving beyond marketing claims. As thyroid ultrasound AI becomes commercial, institutions face procurement decisions. A common failure mode is to select tools based on impressive headline accuracy without understanding validation context. A more robust procurement checklist includes: (1) evidence of external validation on independent datasets. (2) clarity on reference standards and cohort selection; (3) reporting of equipment diversity; (4) description of how the tool handles low-quality or out-of-distribution (inputs unlike the training data) inputs; (5) ability to export structured outputs with model version identifiers; and (6) support for local monitoring and audit. Reviews of available tools suggest that performance differs by product and intended use-case. (2022) highlights that different tools can perform differently and that performance depends on how they are used. Procurement should therefore be aligned with intended use: a reporting assistant has different requirements than a triage tool. Finally, institutions should negotiate update transparency: when the vendor updates the model, what changes, how will it be communicated, and what revalidation support is provided? Without update governance, procurement becomes a one-time decision for a tool that changes over time.

4.11 Cybersecurity and reliability: AI as part of critical clinical infrastructure. AI systems for imaging increasingly rely on networked workflows, cloud inference, or vendor-managed services. This introduces cybersecurity and reliability considerations that are not captured by accuracy metrics. A system that occasionally fails or times out can create workflow disruption and can encourage users to bypass safeguards. Institutions should treat AI inference pipelines like other clinical IT systems: define uptime requirements, fallback behavior (what happens when AI is unavailable), and incident response procedures. If AI outputs are integrated into the report, the system must ensure data integrity—outputs should be traceable to a specific model version and input study. Cybersecurity is also linked to privacy: cine/video data may be more identifiable than static images, and transfers must be secured. Although these issues may feel "outside" endocrinology, they determine whether AI can be adopted safely at scale. In this sense, thyroid ultrasound AI is also a health-IT governance project, not merely an imaging innovation.

4.12 Explainability evaluation: what clinicians actually need to see. Explainability is often framed as heatmaps, but clinicians typically need more practical interpretability: which TI-RADS features drove the suggestion, what uncertainty exists, and what alternative classifications are plausible. Feature-aligned outputs (e.g., composition-first workflows such as Chen et al., 2024) can provide this. For classic CAD and deep learning systems, an effective interface might show: suggested TI-RADS descriptors, a confidence level per descriptor, and a combined category with an explanation of which descriptors contributed most. The goal is to support "reasoned agreement" rather than passive acceptance. For multimodal copilots (Yao et al., 2025), explainability extends to text: the system should be able to cite which image cues support its statements, and it should avoid generating definitive claims when evidence is weak. From a research perspective, explainability should be evaluated empirically: does it improve error detection? does it reduce automation bias? does it shorten time-to-decision without sacrificing accuracy? These are measurable outcomes and should become part of study design.

4.13 Patient safety monitoring: auditing missed cancers without creating perverse incentives. Any triage-support tool raises the question of false negatives. A responsible monitoring strategy includes regular audits of malignancies diagnosed after initial benign classification or low-risk recommendation. However,

audits can create perverse incentives if they are interpreted as blame. A better approach treats audits as learning opportunities: categorize failures (ultrasound scanning quality, atypical lesion, model limitation, threshold issue), update protocols accordingly, and communicate lessons in a non-punitive format. In thyroid ultrasound, where many cancers are indolent, defining "missed clinically significant cancer" should be explicit. Monitoring should also track follow-up adherence, because a low-risk recommendation is only safe if follow-up systems work. This again illustrates why AI deployment is a pathway intervention: the outcome depends on systems of care, not only on the model. Institutions that deploy AI without investing in monitoring risk losing trust and may abandon potentially beneficial tools after avoidable incidents.

4.14 Equity-focused evaluation as a routine part of deployment. Equity evaluation is often discussed abstractly, but it can be operationalized. In thyroid ultrasound AI, useful stratifiers include: age, sex, nodule size, and equipment type. If demographic data are imperfect, equipment stratification is still feasible and informative, because domain shift by machine is common. Institutions can monitor whether AI-assisted decisions change biopsy rates differently across groups. If disparities appear, they may reflect model bias, differences in presentation, or workflow differences (who gets scanned, by whom, on what machine). The key is to detect and investigate rather than assume uniform benefit. Meta-analyses that restrict to external datasets (Gatta et al., 2025) represent a methodological step toward generalizability, but equity requires going further: diverse populations and transparent subgroup reporting. Over time, equity monitoring should become part of the AI quality management system, similar to how infection rates or readmission rates are monitored in quality programs.

4.15 Research transparency and reproducibility as public goods. In fast-moving AI fields, reproducibility is often sacrificed for novelty, but in medical AI this undermines trust. For thyroid ultrasound AI, progress depends on transparent reporting of datasets, patient-level splits, reference standards, and external validation so that results can be compared and reproduced. Dataset publications and benchmarking resources can accelerate progress when they are well documented and privacy-aware. Journals and reviewers can support this by expecting clear reporting and by valuing external validation and failure-mode analysis as contributions, not only headline accuracy.

4.16 How to interpret meta-analytic numbers without overconfidence. Meta-analyses provide pooled sensitivity and specificity estimates, but pooled numbers can obscure important variation. For example, heterogeneity can arise from differing reference standards, different inclusion of small nodules, and different prevalence. A practical reading is: meta-analyses can tell us that AI is promising, but they cannot tell a hospital what performance to expect on its machines, in its population, with its operators. Therefore, meta-analytic results should be used to justify local piloting and to inform what performance ranges are plausible, not to skip local validation. Network meta-analyses (Yang et al., 2023) are particularly useful for understanding tradeoffs: different systems sit on different sensitivity–specificity operating points. But the choice among operating points is a normative one—how much risk of missed cancer is acceptable in exchange for fewer biopsies? That decision belongs to clinical governance, not to the model alone.

4.17 Designing future systems: from "diagnosis" to "decision support under uncertainty." The most clinically valuable AI systems may not be the ones that output a single diagnosis, but the ones that help manage uncertainty. In thyroid nodules, uncertainty is often the reality: indeterminate cytology, borderline ultrasound features, patient preference differences, and competing guideline interpretations. A decision-support system can add value by making uncertainty explicit, by highlighting what additional information would reduce uncertainty (e.g., more views, elastography, follow-up interval), and by documenting rationale. Cine/video approaches (Zhou et al., 2025) and copilot approaches (Yao et al., 2025) point toward richer interaction. However, richer interaction also means richer responsibility: the system must be transparent about what it knows and what it does not. Future evaluation should therefore include not only accuracy but also how well the system supports appropriate uncertainty management and shared decision-making.

4.18 Limitations and boundaries of responsible use. Even a highly accurate model is not a substitute for clinical judgment. AI should not be used in isolation, and it should not be assumed to generalize to populations or devices outside its validation. Institutions should define boundaries of use: which machines and presets are covered, which patient ages are covered, which lesion types are covered (e.g., cystic vs solid), and what to do when boundaries are violated. Clear boundaries reduce harm and reduce unrealistic expectations among clinicians and patients. Regulatory intended-use statements provide a baseline, but local boundaries may need to be narrower. For example, if local validation shows reduced performance on a specific older ultrasound machine, the AI might be disabled on that machine until mitigations are in place. Such boundary-setting is part of mature clinical governance and should be considered a success, not a failure.

4.19 Closing synthesis: what the last five years teach us. Over the last five years, the field has moved from optimistic retrospective studies to more realistic evaluations. Retrospective CAD studies support the intuition that AI can standardize interpretation (Kang et al., 2021; Huang et al., 2022). Prospective workflow-oriented evidence shows that operator dependence can blunt benefit (Edström et al., 2025). Multicenter deep learning and cine/video methods aim to reduce frame-selection sensitivity (Chen et al., 2024; Zhou et al., 2025), while multimodal copilots expand the scope of assistance and raise new governance concerns (Yao et al., 2025). The consistent message is that translation is a health-IT and human-factors project as much as it is an algorithmic project. When that sociotechnical framing is adopted, AI becomes easier to evaluate realistically and, potentially, to deploy safely.

4.20 Workflow measurement: time, cognitive load, and error recovery. If AI is meant to improve efficiency, studies should measure time-to-report and cognitive load proxies (e.g., number of clicks, interruptions) alongside accuracy. A tool that improves AUC but adds friction may be rejected by clinicians. Conversely, a tool that slightly improves agreement while reducing documentation time may have substantial practical value. Error recovery is another overlooked metric: when AI is wrong, how easy is it for the user to recognize the error and correct it? Interfaces that show intermediate TI-RADS descriptors can make correction easier than black-box labels, and this should be considered a safety and usability advantage.

4.21 Integrating AI with multidisciplinary pathways. Thyroid nodule management often involves radiology, endocrinology, surgery, and pathology. AI outputs need to be legible and acceptable to all clinical stakeholders. For example, surgeons may prioritize sensitivity, while endocrinologists may prioritize avoiding unnecessary surgery and biopsies. In multidisciplinary tumor boards, AI-generated structured descriptors may be more useful than a single risk score, because they can be discussed and challenged. Deployment planning should therefore involve representatives from all pathway clinical stakeholders and should define how AI outputs are communicated in referrals and notes.

4.22 Follow-up systems as a safety net. Many low-risk decisions rely on follow-up ultrasound. If follow-up systems are weak—missed appointments, poor recall systems—then any shift toward fewer biopsies can increase risk. Therefore, AI deployment should be paired with evaluation of follow-up compliance and with improvements in reminder systems and scheduling workflows. In settings with robust follow-up, AI can safely reduce immediate procedures. In settings with weak follow-up, conservative thresholds may be necessary, or AI may be limited to reporting support rather than triage.

4.23 Bias from ‘ground truth’ itself. Pathology is often treated as definitive, but labels can still be noisy: sampling error in FNA, interobserver variability in cytology, and selection bias in which nodules go to surgery. This matters because AI models learn patterns associated with the labeled subset of nodules. Dataset papers such as Hou et al. (2024) are valuable, but users should recognize that even pathology-labeled datasets represent a particular slice of clinical reality. Robust evaluation therefore benefits from multiple reference signals (pathology when available, follow-up confirmation for benign nodules) and transparent reporting of how labels were established.

4.24 What to do when AI and clinician disagree. Disagreement management is a design choice. Options include: treat AI as advisory, require a second human review when disagreement occurs, or default to the more conservative action (e.g., biopsy). Each option has different workload and safety implications. A balanced approach is to define disagreement as a trigger for additional scrutiny: re-check key features, review cine loops, or consult an experienced reader. Importantly, disagreement rates themselves can be monitored as a quality metric. sudden increases may indicate drift or workflow changes.

4.25 From pilots to sustained practice. Many AI pilots succeed technically but fail operationally because monitoring and ownership fade after the initial enthusiasm. Sustained adoption requires assigning long-term ownership, budgeting for integration and monitoring, and embedding the tool into routine quality meetings. If AI is treated as a one-off innovation project, it will not be maintained. If it is treated as clinical infrastructure—like PACS or structured reporting—it can be governed and improved over time. This infrastructure mindset is especially important for thyroid ultrasound AI because of operator dependence and the need for continuous training.

4.26 Small-nodule policy and shared decision-making. In many guidelines, subcentimeter nodules are not routinely biopsied unless there are high-risk features or clinical risk factors. AI systems specifically targeting small nodules can therefore influence a sensitive area where overdiagnosis is a concern. A responsible deployment should align AI behavior with local policy: the tool can highlight suspicious features but should not automatically escalate to biopsy without considering size thresholds and clinical context. In practice, the

best use may be to support consistent documentation and to facilitate shared decision-making conversations rather than to "hunt" cancers aggressively in tiny nodules.

4.27 Quantifying uncertainty as a workflow signal. Uncertainty can be operationalized as a triage mechanism: high-confidence low-risk cases can proceed through streamlined follow-up, while low-confidence cases trigger additional imaging or expert review. This reframes uncertainty from a weakness to a safety feature. For ultrasound AI, uncertainty can be driven by image quality, atypical appearance, or domain shift. Capturing and logging uncertainty also supports monitoring, because changes in the uncertainty distribution over time can signal drift or changes in ultrasound scanning practice.

4.28 Final practical takeaway. The current evidence supports AI as a promising assistant, but it also shows that real-world value depends on how AI is embedded: ultrasound scanning protocols, interface design, thresholds, and monitoring. Institutions that treat AI as part of a quality management program—rather than as a stand-alone classifier—are more likely to achieve safer adoption and sustained benefit.

4.29 Measuring success at the system level. Ultimately, thyroid ultrasound AI should be judged by whether it improves the health system's ability to deliver consistent, guideline-concordant care, not by whether it wins a benchmark. A practical way to do this is to define a small "AI scorecard" that is reviewed quarterly by a multidisciplinary group (radiology, endocrinology, surgery, pathology, and IT). Useful indicators include: (1) variability in TI-RADS category assignment across operators, (2) proportion of reports with complete structured descriptors, (3) biopsy yield over time (malignancies per biopsies performed), (4) rates of repeat FNA and downstream surgery for benign disease, (5) time from first ultrasound to definitive diagnosis for malignant nodules, and (6) patient-reported anxiety and satisfaction where feasible. Importantly, these should be stratified by equipment type and clinic site to detect domain shift or uneven benefit. If the tool is intended to reduce unnecessary procedures, then the scorecard should also track the number of avoided biopsies and the number of delayed malignancy diagnoses detected on follow-up. These indicators connect AI performance to social outcomes: workload distribution, resource use, and patient experience. They also help avoid the trap of optimizing a single metric (such as AUC) at the expense of pathway health. If a tool improves diagnostic discrimination but worsens documentation, increases follow-up burden, or erodes trust because outputs are difficult to explain, it may not be a net improvement. Framing success in this broader way is consistent with the journal's focus on technology in social systems and clarifies why governance and health-IT integration are as important as model architecture.

4.30 A minimum standard for "ready-to-use" claims. For a thyroid ultrasound AI tool to be credibly called ready for routine use, the evidence base should include external validation, clear reporting of reference standards and data splitting, a description of ultrasound scanning requirements, and at least one evaluation that measures workflow performance or decision-related outcomes. Without these elements, accuracy claims should be treated as preliminary rather than practice-changing.

4.31 Evaluating copilots without confusing fluency for correctness. Multimodal copilots can produce fluent explanations and suggested management language. The main evaluation risk is that readers (and patients) interpret fluency as correctness. Therefore, copilot evaluation should include "faithfulness" checks: are the generated statements grounded in visible image features and in the documented clinical context? This can be tested with structured checklists where experts mark each generated claim as supported, unsupported, or ambiguous. Another important metric is correction behavior: when the copilot makes an error, do users notice and fix it, or do they accept it? This is especially important if copilot text can be inserted into the medical record. Because thyroid ultrasound reports often serve as the basis for referral decisions, uncorrected AI-generated mistakes can propagate downstream. In practice, a safer design is to generate structured draft fields first, then generate narrative text only from clinician-approved fields.

4.32 Data retention and audit trails: how long should AI outputs be kept? Auditing requires retention of outputs and links to inputs. But retention has costs: storage, privacy risk, and governance overhead. A practical compromise is tiered retention: store structured outputs and model version identifiers long-term, while storing full cine loops only when clinically indicated or for a limited period consistent with imaging retention policies. If cine/video AI is used, institutions should decide whether raw cine loops used for inference are stored as part of the clinical record or as a separate audit artifact. Clear policies reduce ambiguity and help institutions respond to incident investigations. Retention policy should also anticipate vendor updates: if a model changes, historical outputs should remain interpretable in terms of the version that produced them.

4.33 Standardization across institutions: why multicenter collaboration matters. One reason thyroid ultrasound AI evidence is heterogeneous is that ultrasound scanning and reporting practices vary. Multicenter collaborations can harmonize protocols—what views are required, how descriptors are defined, how

indeterminate cases are handled—and can create shared evaluation sets. This reduces duplication and enables more meaningful comparisons between tools. It also supports equitable progress: smaller centers can contribute data under shared governance and benefit from common infrastructure rather than building isolated datasets. In the long run, standardization is as much a social agreement as it is a technical protocol, which is why it fits naturally within a technology-and-society lens.

4.34 Public health perspective: balancing overdiagnosis and timely cancer detection. Thyroid cancer epidemiology includes a well-known risk of overdiagnosis driven by imaging. AI could either worsen or improve this balance. If AI makes clinicians more likely to label nodules as suspicious, it could increase biopsies and diagnoses of indolent cancers. If AI supports more confident low-risk classification, it could reduce unnecessary interventions. Therefore, the public health impact depends on how AI is tuned and governed. Deployment should be accompanied by monitoring of biopsy rates, surgical rates, and the stage distribution of detected cancers. These outcomes connect algorithmic choices to societal consequences—costs, morbidity, and anxiety—making them essential for responsible evaluation.

4.35 Sustainability and compute: a rarely discussed implementation constraint. Cine/video models and multimodal copilots can increase compute requirements. In large health systems, compute translates into cost and environmental footprint. While this may seem peripheral, it affects scalability and may influence which tools are adopted. Efficient inference, edge deployment strategies, and selective use (e.g., running heavy models only on low-confidence cases) can reduce compute burden without sacrificing safety. Sustainability considerations also intersect with equity: expensive compute requirements can limit adoption to high-resource settings.

4.36 Ethical framing: autonomy, transparency, and respect for uncertainty. The ethical goal of AI in thyroid ultrasound is not to create an illusion of certainty, but to support better decisions under uncertainty. Respecting patient autonomy means providing understandable rationale for recommendations, acknowledging uncertainty, and avoiding paternalistic "the AI says…" framing. Transparency also applies to clinicians: they should know what data the model was trained on, what it has been validated on, and where it is likely to fail. Ethical deployment therefore overlaps with good engineering: uncertainty indicators, clear boundaries of use, and monitoring all support both safety and ethical practice.

4.37 Quality improvement cycles: turning AI into a learning loop. A mature deployment treats AI as part of a continuous quality improvement (CQI) cycle. The cycle begins with baseline measurement: existing biopsy yield, reporting completeness, and interobserver variation. After deploying AI in a limited pilot, the same metrics are remeasured, and disagreements are reviewed in a structured way. This review should include both "AI wrong" and "human wrong" cases, because the goal is not to score points but to understand failure modes and improve the pathway. Over time, CQI can refine thresholds, improve ultrasound scanning protocols, and tailor training. Importantly, CQI requires data infrastructure: structured AI outputs, linkage to outcomes, and dashboards that clinical stakeholders trust. Without CQI, AI remains a static tool whose performance is assumed rather than demonstrated in the local context.

4.38 Cross-disciplinary acceptance: aligning incentives and language. Thyroid ultrasound sits at the intersection of specialties. Radiologists may focus on imaging descriptors, endocrinologists on longitudinal risk and patient preference, surgeons on operability and risk of malignancy, and pathologists on cytology categories. AI tools that output a single "benign/malignant" label can clash with these different professional lenses. Tools that align with TI-RADS descriptors and produce structured outputs are easier to integrate because they speak a common language. Adoption planning should explicitly address incentives: if AI reduces biopsies, pathology revenue and workload may change. if it increases structured reporting, radiology workload may shift. Transparent discussion of these impacts can prevent resistance and foster shared ownership.

4.39 Validation in the presence of prevalence shift. Diagnostic accuracy depends on disease prevalence, but in thyroid nodules prevalence varies widely depending on whether a cohort is general screening, referred for FNA, or selected for surgery. Many AI studies are trained and evaluated on high-prevalence cohorts, which can inflate positive predictive value in the reported setting but can mislead clinicians in low-prevalence environments. Local validation should therefore estimate predictive values under local prevalence and should consider how AI would behave when most nodules are benign. This is also where calibration matters: a model that ranks well may still produce probabilities that are inappropriate for a lower-prevalence setting unless recalibrated.

4.40 Handling drift from workflow change. Drift is not only a model issue. it can be a workflow issue. If a hospital changes its ultrasound preset, hires new sonographers, or changes its referral criteria, the distribution of images and nodules changes. AI monitoring should therefore be integrated with operational

change management. When a change occurs, the AI steward should trigger a "revalidation checkpoint": a small sample review to ensure outputs remain sensible. This is analogous to how labs validate new reagent lots or instruments. Treating workflow change as a drift trigger makes monitoring proactive rather than reactive.

4.41 Ethical use in patient-facing contexts. Some institutions may be tempted to show AI outputs directly to patients as part of transparency or patient portals. This can be beneficial if the outputs are well-calibrated and explained, but risky if patients interpret AI scores as definitive diagnoses. If patient-facing use is considered, outputs should be accompanied by clear explanations, uncertainty statements, and guidance on next steps. Alternatively, institutions can keep AI outputs clinician-facing but provide general explanations to patients about the role of AI in standardizing assessment. The choice should be deliberate and consistent with local patient communication practices.

4.42 Convergence with other thyroid imaging modalities and future integration. Ultrasound AI is increasingly discussed alongside elastography, contrast-enhanced ultrasound, and even molecular testing in indeterminate cytology. While this review focuses on ultrasound, future decision support may integrate multiple modalities. Multimodal architectures (Yao et al., 2025) point in this direction. For health-IT planning, this suggests that the infrastructure built for ultrasound AI—structured outputs, versioning, outcome linkage—can serve as a foundation for broader endocrine decision support systems.

Beyond ultrasound, other imaging modalities are also being explored for AI-assisted thyroid nodule assessment in preoperative settings. For example, a recent systematic review and meta-analysis evaluated CT-based radiomics and deep-learning models for preoperative thyroid nodule classification and included comparisons with radiologists, reporting promising pooled performance while emphasizing heterogeneity and validation quality (Broomand Lomer et al., 2026).

4.43 Final note on responsible enthusiasm. AI in thyroid ultrasound is moving quickly, and it is easy to be either overly optimistic or overly skeptical. The evidence supports cautious optimism: multiple studies and syntheses show strong diagnostic discrimination, but real-world impact depends on operator behavior, workflow integration, and governance. The most productive stance is to treat AI as a tool that can improve consistency when embedded in a quality system. When institutions invest in ultrasound scanning standards, structured reporting, monitoring, and multidisciplinary oversight, AI becomes more likely to deliver value. When those sociotechnical supports are absent, even a strong model can fail to improve care.

5. Conclusions

AI for thyroid nodule ultrasound is moving from retrospective proof-of-concept toward workflow-oriented validation, cine/video systems, and multimodal copilots. Evidence suggests potential for standardization and support for less experienced users, but prospective workflow studies highlight strong operator dependence and variable real-world impact (Edström et al., 2025).

Safe adoption requires external and local validation, calibration and uncertainty communication, structured reporting integration, and governance for monitoring, accountability, and equitable performance over time.

Conflicts of Interest: No conflicts of interest to declare.

REFERENCES

1. Broomand Lomer, N., Ahmadzadeh, A. M., Ashoobi, M. A., Abdi, S., Ghasemi, A., & Gholamrezanezhad, A. (2026). CT-based radiomics and deep learning for preoperative thyroid nodule classification: A systematic review, meta-analysis, and radiologist comparison. *Academic Radiology, 33*(1), 98–116. https://doi.org/10.1016/j.acra.2025.09.045
2. Ni, J., You, Y., Wu, X., Chen, X., Wang, J., & Li, Y. (2025). Performance evaluation of deep learning for the detection and segmentation of thyroid nodules: Systematic review and meta-analysis. *Journal of Medical Internet Research, 27*, e73516. https://doi.org/10.2196/73516
3. Potipimpanon, P., Charakorn, N., & Hirunwiwatkul, P. (2022). A comparison of artificial intelligence versus radiologists in the diagnosis of thyroid nodules using ultrasonography: A systematic review and meta-analysis. *European Archives of Oto-Rhino-Laryngology, 279*(11), 5363–5373. https://doi.org/10.1007/s00405-022-07436-1
4. Xu, J., Xu, H.-L., Cao, Y.-N., Huang, Y., Gao, S., Wu, Q.-J., & Gong, T.-T. (2023). The performance of deep learning on thyroid nodule imaging predicts thyroid cancer: A systematic review and meta-analysis of epidemiological studies with independent external test sets. *Diabetes & Metabolic Syndrome, 17*(11), 102891. https://doi.org/10.1016/j.dsx.2023.102891

5. Chen, C., Jiang, Y., Yao, J., Lai, M., Liu, Y., Jiang, X., Ou, D., Feng, B., Zhou, L., Xu, J., Wu, L., Zhou, Y., Yue, W., Dong, F., & Xu, D. (2024). Deep learning to assist composition classification and thyroid solid nodule diagnosis: A multicenter diagnostic study. *European Radiology, 34*(4), 2323–2333. https://doi.org/10.1007/s00330-023-10269-z
6. Edström, A. B., Makouei, F., Wennervaldt, K., Lomholt, A. F., Kaltoft, M., Melchiors, J., Hvilsom, G. B., Bech, M., Tolsgaard, M., & Todsen, T. (2025). Human–AI collaboration for ultrasound diagnosis of thyroid nodules: A clinical trial. *European Archives of Oto-Rhino-Laryngology, 282*(6), 3221–3231. https://doi.org/10.1007/s00405-025-09236-9
7. Gatta, E., Gatta, R., Morandi, R., Isoli, S., Corvaglia, S., Vetrugno, S., Maltese, V., Pirola, I., Casella, C., & Cappelli, C. (2025). Machine learning for diagnosis of malignant thyroid nodules based on thyroid ultrasound: Systematic review and meta-analysis of studies with external datasets. *European Journal of Radiology Open*. Advance online publication. https://doi.org/10.1016/j.ejro.2025.100716
8. Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E., Pacini, F., Randolph, G. W., Sawka, A. M., Schlumberger, M., Schuff, K. G., Sherman, S. I., Sosa, J. A., Steward, D. L., Tuttle, R. M., & Wartofsky, L. (2016). 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. *Thyroid, 26*(1), 1–133. https://doi.org/10.1089/thy.2015.0020
9. Hou, X., Hua, M., Zhang, W., Ji, J., Zhang, X., Jiang, H., Li, M., Wu, X., Zhao, W., Sun, S., Cao, L., & Wang, L. (2024). An ultrasonography of thyroid nodules dataset with pathological diagnosis annotation for deep learning. *Scientific Data, 11*(1), 1272. https://doi.org/10.1038/s41597-024-04156-5
10. Huang, P., Zheng, B., Li, M., Xu, L., Rabbani, S., Mayet, A. M., Chen, C., Zhan, B., & Jun, H. (2022). The diagnostic value of artificial intelligence ultrasound S-Detect technology for thyroid nodules. *Computational Intelligence and Neuroscience, 2022*, 3656572. https://doi.org/10.1155/2022/3656572
11. Kang, S., Lee, E., Chung, C. W., Jang, H. N., Moon, J. H., Shin, Y., Kim, K., Li, Y., Shin, S. M., Kim, Y. H., Kwon, S. K., Ahn, C. H., Jung, K. Y., Hong, A. R., Park, Y. J., Park, D. J., Kwak, J. Y., & Cho, S. W. (2021). A beneficial role of computer-aided diagnosis system for less experienced physicians in the diagnosis of thyroid nodule on ultrasound. *Scientific Reports, 11*, 20448. https://doi.org/10.1038/s41598-021-99983-6
12. Russ, G., Bonnema, S. J., Erdogan, M. F., Durante, C., Ngu, R., & Leenhardt, L. (2017). European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: The EU-TIRADS. *European Thyroid Journal, 6*(5), 225–237. https://doi.org/10.1159/000478927
13. See-Mode Technologies Pte. Ltd. (2024). *510(k) summary: See-Mode Augmented Reporting Tool, Thyroid (SMART-T) (K240697)*. U.S. Food and Drug Administration. https://www.accessdata.fda.gov/cdrh_docs/pdf24/K240697.pdf
14. 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., Hammers, L. W., Hamper, U. M., Langer, J. E., Reading, C. C., Scoutt, L. M., & Stavros, A. T. (2017). ACR thyroid imaging, reporting and data system (TI-RADS): White paper of the ACR TI-RADS committee. *Journal of the American College of Radiology, 14*(5), 587–595. https://doi.org/10.1016/j.jacr.2017.01.046
15. Vahdati, S., Khosravi, B., Robinson, K. A., Rouzrokh, P., Moassefi, M., Akkus, Z., & Erickson, B. J. (2024). A multi-view deep learning model for thyroid nodules detection and characterization in ultrasound imaging. *Bioengineering, 11*(7), 648. https://doi.org/10.3390/bioengineering11070648
16. Wildman-Tobriner, B., Taghi-Zadeh, E., & Hoang, J. K. (2022). Artificial intelligence (AI) tools for thyroid nodules on ultrasound, from the AJR Special Series on AI applications. *AJR American Journal of Roentgenology, 219*(4), 1–8. https://doi.org/10.2214/AJR.22.27430
17. Yang, L., Li, C., Chen, Z., He, S., Wang, Z., & Liu, J. (2023). Diagnostic efficiency among EU-/C-/ACR-TIRADS and S-Detect for thyroid nodules: A systematic review and network meta-analysis. *Frontiers in Endocrinology, 14*, 1227339. https://doi.org/10.3389/fendo.2023.1227339
18. Yao, J., Wang, Y., Lei, Z., Wang, K., Feng, N., Dong, F., Zhou, J., Li, X., Hao, X., Shen, J., Zhao, S., Gao, Y., Wang, V., Ou, D., Li, W., Lu, Y., Chen, L., Yang, C., Wang, L., Feng, B., Zhou, Y., Chen, C., Yan, Y., Wang, Z., Ru, R., Chen, Y., Zhang, Y., & Xu, D. (2025). Multimodal GPT model for assisting thyroid nodule diagnosis and management. *npj Digital Medicine, 8*, 245. https://doi.org/10.1038/s41746-025-01652-9
19. Zhan, J., Zhang, J., Zhu, S., Ni, L., Zhang, C., & Hu, J. (2025). Diagnostic performance of ultrasound characteristics-based artificial intelligence models for thyroid nodules: A systematic review and meta-analysis. *Frontiers in Oncology, 15*, 1614603. https://doi.org/10.3389/fonc.2025.1614603
20. Zhou, Y., Chen, C., Yao, J., Yu, J., Feng, B., Sui, L., Yan, Y., Chen, X., Liu, Y., Zhang, X., Wang, H., Pan, Q., Zou, W., Zhang, Q., Lin, L., Xu, C., Yuan, S., He, Q., Ding, X., Liang, P., Wang, V. Y., & Xu, D. (2025). A deep learning based ultrasound diagnostic tool driven by 3D visualization of thyroid nodules. *npj Digital Medicine, 8*, 126. https://doi.org/10.1038/s41746-025-01455-y