

**ARTIFICIAL INTELLIGENCE IN GYNECOLOGICAL ULTRASONOGRAPHY:  
DIAGNOSTIC CAPABILITIES, CLINICAL VALIDATION, AND LIMITATIONS**

*Xolboyeva Mashhura Yoqutjon qizi*

Email – [yoqutjonovnam@gmail.com](mailto:yoqutjonovnam@gmail.com)

*Termez University of Economics and Service*

*Faculty of Medicine*

**Abstract**

**Background:** Artificial intelligence (AI) is increasingly proposed as an adjunct to gynecological ultrasonography, including assessment of adnexal masses, endometrial lesions, adenomyosis, uterine leiomyomas, and reproductive imaging. However, model discrimination reported in curated datasets does not automatically establish safety, calibration, transportability, or patient benefit. **Objective:** To critically evaluate diagnostic opportunities, methodological weaknesses, real-world evidence pathways, and regulatory considerations relevant to AI-assisted gynecological ultrasound. **Methods:** A focused evidence-informed narrative synthesis considered representative clinical accuracy studies, systematic reviews, reporting standards, and guidance from professional and regulatory bodies through October 2026. The manuscript differentiates published results from proposed future analyses; no de novo database search exhaustiveness, patient recruitment, or pooled meta-analysis is claimed. **Results:** A 2025 systematic review of 44 studies concerning AI-assisted diagnosis of gynecological pelvic tumors found a pooled ovarian-tumor discrimination area under the curve of 0.89 (95% confidence interval 0.85–0.92), but approximately 95% of included studies were judged at high risk of bias; external testing and calibration were uncommon. The externally validated ADNEX clinical model offers a necessary comparator rather than an obsolete benchmark. Across use cases, image-level splitting, spectrum restriction, operator and device shift, imperfect reference standards, and missing clinical covariates can inflate apparent model performance. Real-world deployment requires prospective validation, clinically meaningful workflow endpoints, oversight, and postmarket monitoring. **Conclusion:** AI should currently be evaluated as calibrated, auditable decision support within defined diagnostic pathways, not as an autonomous replacement for expert ultrasonography. Multicentre patient-level validation, subgroup fairness analyses, transparent reporting, and prospective impact studies are prerequisites for credible adoption. For Uzbekistan,

a staged national data-quality and governance programme would be more defensible than immediate large-scale automation.

**Keywords:** gynecological ultrasonography; artificial intelligence; ovarian neoplasms; adnexal masses; endometrial pathology; adenomyosis; deep learning; machine learning; clinical validation; diagnostic accuracy; real-world evidence; algorithmic bias; explainable AI; decision support.

## 1. INTRODUCTION

### Clinical scope

Gynecological ultrasonography is a dynamic, operator-dependent examination in which diagnostic value arises from more than pixels. The examiner integrates age, menstrual status, symptoms, prior imaging, Doppler patterns, palpation, probe pressure, and the evolution of appearances during the examination. Any proposed artificial-intelligence system inherits this clinical context whether or not its training dataset actually records it. AI might standardize measurements, highlight overlooked features, stratify adnexal malignancy risk, or support audit; it may also amplify errors by making a superficially plausible prediction from a narrow, selected image. This distinction motivates evaluation at the patient and decision level rather than only at the frame level. [1–4]

### Diagnostic evidence landscape

A contemporary systematic review and meta-analysis examined 44 studies of ultrasound AI for gynecologic pelvic tumors, comprising predominantly ovarian lesions alongside limited endometrial and myometrial evidence. Its ovarian pooled AUC was 0.89 (95% CI 0.85–0.92); approximately 95% of included investigations had high risk of bias, and very few assessed external validity or calibration. These figures support technical promise but expose a substantial translational gap. The statistical estimate describes reported study populations and tasks; it does not prove superior patient outcomes or justify universal deployment. Established IOTA rules and ADNEX constitute clinically interpretable comparators. [1–3]

### RWD and RWE

Real-world data (RWD) are information collected during care rather than exclusively for a tightly controlled efficacy experiment: image archives, electronic health records, pathology, referral outcomes, prescriptions, registries, insurance claims, and patient-reported results. Real-world evidence (RWE) is clinical evidence generated through an appropriate design and analysis of such data. A large archive is not automatically evidence; exposure definitions, indication, completeness, verification, confounding,

and provenance determine its interpretability. In ultrasound AI, a dataset of labeled scans without follow-up may be inadequate to estimate safety among women not referred to surgery. [10–12]

### RCT versus real-world comparison

Randomized controlled trials (RCTs) can estimate whether clinician access to an AI tool changes management, referral, time to treatment, missed malignancy, or resource use under a specified protocol. Observational RWE more readily captures varied scanners, operators, regions, socioeconomic circumstances, pregnancy status, and evolving workflow, but is vulnerable to confounding and selection bias. The appropriate comparison is not “RCT or RWE”; model development, diagnostic validation, randomized clinical impact evaluation, and postdeployment surveillance answer different questions. [10,12,16]

### Efficacy and effectiveness

Algorithmic efficacy in a controlled benchmark, even with strong AUC or accuracy, differs from clinical effectiveness. Effectiveness means an improvement in patient-relevant outcomes when a tool is used within actual referral patterns, staffing levels, reporting systems, and accountability arrangements. For example, an AI system may classify stored images correctly yet have limited effectiveness if endometrial biopsies are delayed, pathology cannot be linked, clinicians distrust uncertainty estimates, or unsafe false reassurance results in missed follow-up. The gap between discrimination and implementation resembles the distinction between drug efficacy in explanatory trials and effectiveness in pragmatic care. [1,10,16]

### WHO FDA EMA roles

WHO emphasizes autonomy, inclusion, equity, transparency, public accountability and safety in medical AI. The FDA regulates intended medical-device functions rather than granting blanket approval to “AI”; its Good Machine Learning Practice framework and real-world-evidence guidance emphasize lifecycle management, representative datasets and fit-for-purpose evidence. EMA’s DARWIN EU infrastructure concerns regulatory real-world evidence for medicines, not a certification pathway for ultrasound imaging software. These differences matter when a manuscript spans drugs, devices and digital care. [10–15]

## 2. MATERIALS AND METHODS

### Review design and boundaries

This article is a critical integrative narrative review, not a registered systematic review or original diagnostic trial. Representative studies, consensus frameworks, systematic

reviews and regulatory documents were sought through accessible literature indexing and official publications. PubMed-indexed material, ultrasound specialty literature, BMJ, Nature Medicine, WHO, FDA and EMA publications were emphasized; Scopus, Web of Science, Embase and Cochrane Library are important target databases for any future reproducible systematic search, but exhaustive individual searches of these subscription resources are not asserted here. A future PRISMA review should register its protocol, retain complete dates and strings, screen independent duplicates, record exclusions, and appraise risk of bias.

### Eligibility and synthesis

Directly relevant works covered AI-assisted transvaginal or transabdominal gynecologic ultrasound, clinical diagnostic standards such as IOTA/ADNEX and O-RADS, image-based prediction, reporting guidance, regulatory policy and real-world research. Priority was assigned to 2020–2026 sources while landmark earlier measurement and classification studies were retained. Clinical and methodologic findings were narrated by target condition, reference standard, performance outcome, validation design and applicability. No invented odds ratio, hazard ratio, synthetic patient cohort, meta-analytic heterogeneity statistic or pooled sensitivity was calculated for this manuscript.

### Quality framework

A future systematic assessment should use QUADAS-2 or QUADAS-AI for diagnostic test studies; PROBAST or PROBAST+AI for prediction models where applicable; STARD-AI for diagnostic accuracy reporting; TRIPOD+AI for prediction-model reporting; CLAIM for imaging-specific transparency; and CONSORT-AI or SPIRIT-AI when prospective interventional evaluation is undertaken. The unit of split must be the patient, preferably across sites and time, because images from the same examination are highly correlated. Data leakage can arise through machine identifiers, near-duplicate stills or correlated pathology labels. [6–9]

### Analytical estimands

Sensitivity, specificity, likelihood ratios, calibration intercept and slope, AUC, decision-curve net benefit, time to diagnosis and false-reassurance frequency are noninterchangeable endpoints. The threshold should be prespecified according to downstream referral capacity and the relative harm of a missed cancer versus unnecessary surgery. For multiclass ovarian models, marginal category performance matters because aggregate AUC can conceal poor discrimination in borderline tumors or rare histologies. Uncertainty intervals should use patient-level resampling, with centre-level clustering where appropriate.

### Causal-inference toolkit

Propensity-score matching could construct comparable patient groups in observational comparisons of AI-assisted versus conventional ultrasonography; inverse probability weighting could estimate average treatment-policy effects under exchangeability and positivity assumptions. A target-trial emulation would specify eligibility, assignment strategies, time zero, follow-up, outcome and estimand before using routine-care records. Cox regression and Kaplan–Meier methods could characterize time to delayed diagnosis, surgery or recurrence, but only with actual longitudinal follow-up, censoring assumptions and proportional-hazards checks. None of these procedures was run in this review.

### ML, DL and language methods

Machine learning approaches include penalized regression, gradient boosting, random forests and support-vector machines on structured morphology features. Deep learning approaches include convolutional neural networks, transformers, segmentation architectures and multimodal fusion of still images, cine clips and structured clinical variables. Natural language processing can extract laterality, lesion morphology and recommendations from free-text reports; large language models may support summarization but require source linking, human verification and privacy controls. Reinforcement learning for scan acquisition or examination sequencing remains conceptually interesting but lacks broad validated autonomous gynecological diagnostic deployment.

## 3. RESULTS: EVIDENCE SYNTHESIS

### Ovarian mass malignancy discrimination

The most mature gynecological ultrasound AI literature addresses adnexal masses because surgery and histology provide relatively concrete labels for selected patients. Learned texture and morphology can capture papillary projections, septations, solid components, vascular patterns and acoustic shadowing, but incidental associations may dominate a model trained in one referral centre. Pooled discrimination is encouraging, yet interpretation must be tempered by the systematic review's high risk-of-bias judgments. [1]

IOTA Simple Rules and ADNEX are indispensable comparators. A valid study should evaluate all three on exactly the same consecutive participants, document indeterminate rule outcomes, compare calibration and referral decision thresholds, and involve experienced and less experienced operators. An apparent performance advantage over an artificially weak baseline is not a demonstration of clinical added value. [2,3]

Training solely on operated masses creates verification and spectrum bias because typical low-risk cysts remain unoperated. Histology is unavailable for many benign follow-up cases; these need predefined longitudinal outcome adjudication rather than forced malignancy labels. External multicentre testing, rare-histology subgroup analysis, and prospective decision impact remain priorities.

### O-RADS and standardized lexicons

O-RADS US v2022 provides structured terminology, risk categories and management recommendations for ovarian-adnexal imaging; it explicitly recognizes sonographic features that favor benignity and MRI as an adjunct for selected uncertain lesions. The taxonomy offers both human-readable labels and a quality-control target for software developers. [4]

An AI system could first produce a structured description, then map features into O-RADS categories with clinician review, rather than bypassing the lexicon. This separates image perception from management recommendation and permits auditing whether disagreement occurs at feature extraction or risk calibration. Outputs must communicate when the image is insufficient for reliable classification.

The label “O-RADS category” is not a ground-truth cancer diagnosis. When the model learns human-assigned categories, agreement may simply indicate imitation of raters, while underlying pathology is a different endpoint. Categories can evolve over time, requiring versioned labels and updated evaluation.

### Endometrial cancer and hyperplasia

Endometrial lesion assessment combines endometrial thickness, echogenicity, continuity of the endometrial–myometrial interface, focal lesions, blood-flow architecture and menstrual or menopausal context. A 2025 systematic review identified 30 AI-ultrasound studies related to endometrial cancer, spanning diagnosis and downstream prognostic tasks. These tasks should not be pooled as if they measured a single endpoint. [5]

IETA terminology offers a common framework for extracting endometrial image characteristics. Clinically useful systems must define whether the output predicts biopsy-confirmed disease, triages abnormal uterine bleeding, or estimates stage after confirmed malignancy. Menopausal status, hormone exposure and bleeding pattern modify baseline risk and are necessary to assess calibration. [18]

Endometrial histology is preferentially sampled in symptomatic, high-risk patients. Models based on selected biopsies may not generalize to asymptomatic populations.

Endometrial thickness thresholds also behave differently before and after menopause; pooling these groups without interaction modeling risks misclassification.

### **Adenomyosis**

Adenomyosis produces heterogeneous myometrial morphology, including myometrial cysts, asymmetry, fan-shaped shadowing and junctional-zone abnormalities. It coexists with leiomyomas and endometriosis, making reference-standard definitions and focal-versus-diffuse classification central to model evaluation. A deep model trained on a single static slice may not represent the full organ and may underperform when focal change is subtle.

AI segmentation can potentially standardize myometrial layers and quantify lesion extent, but reference labels should be assigned using agreed sonographic criteria, with MRI or histopathology when available and appropriately indicated. Reader studies should compare expert visual assessment, intermediate readers, and readers assisted by software across the same challenging cases.

A frequent danger is circular validation: two experts annotate the image, the algorithm reproduces their morphology label, and this is then described as confirmed disease. A reference-standard hierarchy, disagreements resolved independently, and blinded reader studies are essential.

### **Leiomyoma measurement and characterization**

Uterine leiomyomas are common candidates for automated segmentation because their size and location are repeatedly documented over time. Longitudinal volume estimates could help follow response to therapy and identify growth patterns, although heterogeneous echogenicity, calcification and shadowing challenge automated contours. A model for presence detection cannot automatically infer molecular subtype or malignancy risk.

Structured lesion mapping with a FIGO-compatible location vocabulary and time-linked measurements could reduce inconsistent documentation between examinations. Clinicians need lesion-level provenance and the ability to correct contours. A useful assessment compares absolute volume error, interobserver reproducibility and clinical decisions rather than reporting only image overlap.

Ground truth for volume itself is approximate because 3D reconstruction, ellipsoid formulas and probe pressure generate different estimates. Overconfident “precise” growth predictions may trigger unnecessary surgery; uncertainty ranges and minimum clinically meaningful change should be prespecified.

### Deep endometriosis and lesion localization

Ultrasound for deep endometriosis involves guided dynamic maneuvers, compartment-by-compartment mapping, mobility tests and assessment of tenderness. These features are incompletely represented by static archived images. AI for automatic recognition may assist learning and triage, but must be developed for a clearly defined anatomic target rather than an undifferentiated label of endometriosis.

Cine clips with temporal registration, standardized scanning protocols and surgical or expert multimodal reference standards could enable validation of compartment-level predictions. Evaluation should stratify bladder, bowel, uterosacral and ovarian disease and avoid considering a normal scan as excluding superficial disease.

Referral-centre enrichment and verification by surgery inflate case severity. A model tested only where expert scanners operate may fail in primary care; transportability must be assessed by location, machine and examination protocol.

### Ovarian reserve and follicle counting

Automated follicle detection and antral follicle counting may reduce repetitive operator work in fertility care, but the clinically relevant target is the ovarian response or a management decision, not simply agreement on drawn circles. Follicle count depends on scanning plane, resolution, menstrual timing and ovarian stimulation context.

A credible evaluation compares assisted and unassisted counting within a multicentre reproducibility protocol, separately assessing polycystic morphology, enlarged ovaries, post-surgical anatomy and low reserve. Prediction of oocyte yield must be calibrated and benchmarked against established AMH and clinical age information.

It would be scientifically unsound to portray a flawless follicle count as a validated predictor of natural conception or oocyte quality. Label disagreement, ovarian visibility and scanner settings should be carried forward into uncertainty estimates.

### Pelvic inflammatory and infectious differential diagnosis

Inflammatory adnexal conditions, tubo-ovarian abscess, hydrosalpinx and hemorrhagic lesions can mimic neoplastic features. Deep models may inadvertently classify images by referral workflow rather than actual morphology, especially when diagnosis is inferred from antibiotics, surgery or report keywords.

Models intended for a mass-triage pathway need out-of-distribution detection and explicit escalation instructions for suspected infection, pregnancy-related emergencies or torsion. Clinical symptoms and blood tests cannot be silently replaced by image-based risk scoring.

The harm from a confident false benign prediction in a time-sensitive infection is different from the harm of unnecessary follow-up. Standard ROC summaries do not capture time-to-treatment effects or the need for urgent action.

### **Torsion and time-critical disease**

Ovarian torsion is an episodic emergency with variable Doppler findings; preserved flow does not reliably exclude it. Training datasets are rare, urgent and highly selected, which is an intrinsically difficult setting for AI development.

Safe tools should prioritize escalation prompts when clinical suspicion is high and avoid autonomous rule-out language. A prospective study might measure time to surgical evaluation, incorrect reassurance and clinician override rather than only retrospective image classification.

Scarce positive cases, intermittent presentations and changing surgical practice make high retrospective accuracy especially fragile. Synthetic augmentation does not manufacture clinical diversity.

### **Transabdominal versus transvaginal protocols**

Transvaginal and transabdominal scans differ in field of view, resolution, depth and patient suitability. Images produced for fetal or abdominal applications should not be assumed to transfer to pelvic pathology tasks.

An implementation study should report probe type, transducer frequency, acquisition protocol, bladder status and machine software. Models may require separate calibration or distinct intended uses by modality.

Combining modalities without metadata can create spurious shortcuts and conceal failure in clinically important groups where transvaginal examination is unavailable or declined.

### **Spectral and color Doppler**

Vascularity supports characterization of solid components and some endometrial lesions, yet Doppler settings are highly machine- and operator-dependent. Color gain, pulse repetition frequency, insonation angle and motion produce patterns that software may misinterpret as biologic signal.

A clinically defensible multimodal network should incorporate acquisition metadata and perform controlled evaluation with and without Doppler. The contribution of each channel should be validated against external clinical sites, not merely optimized in a held-out subset from the same scanner.

Doppler channel dropout, artifacts and manufacturer-specific color maps risk a misleading model confidence. Ablation studies and device-specific drift monitoring are necessary.

### **Radiomics versus end-to-end models**

Handcrafted radiomics may offer more interpretable texture descriptors, whereas end-to-end deep networks learn high-dimensional spatial representations. In published pooled analyses, the apparent hierarchy varies according to study quality, sample size, augmentation and external testing. [1]

Interpretability cannot be equated with truth: a named texture variable can reflect gain rather than disease, while a network attention map may indicate relevance without causal explanation. Both model families need prespecified clinical endpoints and independent validation.

Radiomic features are sensitive to segmentation, normalization and acquisition settings. Deep networks may overfit center signatures. Transparent preprocessing and equipment-stratified reporting can reduce these threats.

### **Calibration and net benefit**

An AUC of 0.89 expresses rank discrimination over case-control pairs; it does not specify whether a reported 20% malignancy risk corresponds to 20% disease frequency among similar patients. Calibration, decision thresholds and net benefit are essential for safely routing patients. [1,3]

Calibration plots, observed-to-expected ratios, calibration slope, and decision-curve analysis should be reported by site and subgroup. Referral thresholds should align with surgical capacity, availability of specialist review and acceptable false-negative harm.

Prevalence shift is common between tertiary oncology and primary gynecology clinics. A model can retain moderate discrimination but become severely miscalibrated, particularly when deployed in lower prevalence populations.

### **Patient-level and temporal validation**

Multiple frames from the same cine loop share structure. Randomly dividing frames between training and testing allows an algorithm to recognize the patient or scanner rather than disease. The 2025 evidence synthesis identifies lack of patient-level separation as a major recurrent concern. [1]

Minimum credible design uses patient-separated internal testing, time-separated assessment and external evaluation at clinics unseen in training. Cross-validation folds must be grouped by patient and preferably by center when testing transportability.

Reporting only accuracy after a convenient split obscures optimistic selection of thresholds and hyperparameters. All model-selection decisions should remain inside the development data.

### [Reader assistance versus replacement](#)

Software may help a novice recognize high-risk morphology or prompt a standard description, but its impact differs from stand-alone diagnostic performance. Reader assistance can create automation bias as well as reduce perceptual misses.

Randomized crossover reader studies should measure accuracy, reading time, confidence, discordance resolution, referral recommendation and unintended over-reliance. Studies should include novice, intermediate and expert readers, with order effects and washout periods considered.

A high stand-alone AI score is neither necessary nor sufficient for net clinician benefit. In difficult cases, an unjustified persuasive confidence number can degrade expert decisions.

### [Missing data and labels](#)

Clinical datasets frequently lack pregnancy status, reliable menstrual history, pathology verification, follow-up and consistent lesion laterality. Missingness is often informative: women without suspicious findings are less likely to undergo surgery.

Audit missingness by center and clinical subgroup; distinguish structurally unavailable data from random omissions; use multiple imputation only when assumptions are plausible. Include missingness indicators with caution and predefine how the deployed tool behaves when essential variables are absent.

Complete-case analysis can intensify referral bias, while automatic imputation can introduce fabricated certainty. Sensitivity analyses for missing-not-at-random mechanisms are preferable to pretending data gaps do not exist.

### [Bias and health equity](#)

Equipment availability, referral selection and clinical documentation may differ by age, rural residence, socioeconomic status, ethnicity and disability. An apparently accurate model can deliver unequal false-negative rates across these groups even if protected attributes were not included as inputs.

Fairness evaluation should include sensitivity, calibration and missed malignancy under clinically justified strata and should use privacy-respecting governance. Representative enrollment and postdeployment auditing are more defensible than solely constraining a training loss function.

Sensitive subgroup data can be incomplete, and fairness definitions conflict. Publish the ethical rationale for chosen comparisons and avoid assertions of equity without relevant measurement.

### **Explainability and uncertainty**

Saliency maps, feature importance and textual rationales can help inspection but do not guarantee a clinically faithful causal explanation. Explanation stability may vary under harmless changes to cropping or image contrast.

Show calibrated uncertainty, missing-input warnings and case retrieval where lawful; allow clinicians to inspect target lesions and correct annotation. Explanation usefulness should be tested prospectively against decision quality rather than assessed by visual appeal.

Large language models may provide fluent but incorrect claims about pathology or evidence. Every clinical explanation must remain linked to verified images, observations and the intended diagnostic endpoint.

### **Distribution shift and drift**

Patient mix, scanner replacement, new reconstruction software, changed coding, shifts in prevalence and ultrasound technician turnover can all modify prediction behavior after deployment.

A deployment registry should log input quality, software version, probe type, rejection rate, subgroup calibration and delayed diagnosis. An escalation and rollback protocol should predefine when drift leads to restriction or suspension of use.

Recalibration itself can become unsafe when fed labels from selective biopsies. Monitoring requires outcome ascertainment and separation between retrospective diagnostic audit and software updates.

### **Resource-limited care**

In resource-limited clinics, improved image capture and structured reporting may bring greater benefit than complicated autonomous inference. Network reliability, electricity, device compatibility and workforce support often determine whether any tool functions safely.

A feasible pilot should measure completeness of pelvic reports, correct referral, exam repeat rates and cost per actionable finding while using transparent clinician-led oversight. A tool that fails offline or on a common lower-cost scanner should not be presented as universally deployable.

Models trained on high-end university hospital equipment can worsen inequality by selectively producing confident outputs for affluent sites. Site-based testing and procurement transparency are vital.

#### 4. REAL-WORLD DATA, AI, SAFETY AND IMPLEMENTATION

##### Electronic medical records and imaging archives

EHRs permit linkage of symptoms, menopausal status, ultrasound episodes, surgical decisions and pathology. Picture archiving and communication systems record images but can have inconsistent metadata. The proper linkage key is a privacy-preserving patient and examination identifier, not a convenience filename. De-identified records require audit of re-identification risk. Longitudinal follow-up helps adjudicate lesions managed conservatively, reducing dependence on surgical cohorts. A transparent data dictionary should reconcile laterality, repeated procedures, incidental findings and changes in diagnosis. [10,12]

##### E-prescriptions and laboratory information

Electronic prescriptions and laboratory systems may capture drug exposure, adverse events, tumor markers and inflammatory investigations, but they are secondary context for imaging diagnosis. Prescription issuance is not adherence, and tumor markers are not perfect reference standards. Time-aligned extraction distinguishes antecedent medication exposure from treatment initiated after a suspicious scan. This temporal ordering is crucial if diagnostic models are later extended to predict treatment response or adverse events. [10–12]

##### Registries, claims and pharmacovigilance

National registries may provide cancer outcomes, surgery, pathology and mortality; claims databases characterize service utilization but may omit morphology and clinical detail. Pharmacovigilance records concern suspected adverse drug reactions and do not prove that imaging AI detected disease. To evaluate AI-assisted management, medication safety endpoints should be specific: inappropriate prescriptions after a wrong diagnosis, contrast-related adverse events when MRI follows uncertain ultrasound, or delayed treatment. Signal detection requires adjudication and denominators; spontaneous reports cannot directly estimate incidence. [11,12]

##### Wearables, mobile apps and patient-reported outcomes

Wearables and menstrual applications can record bleeding, pain, activity and symptom trajectory. Such sources may enrich triage when a patient undergoes ultrasound for abnormal uterine bleeding or chronic pelvic pain. Yet digital engagement is selective and can overrepresent younger, digitally connected users; inferred cycles do not

establish histologic disease. Consent, data minimization, missingness and access to a non-digital alternative are essential to fair multimodal research.

### [Machine learning, deep learning and multimodal analysis](#)

Classical machine learning can combine standardized lesion descriptors and demographics with calibrated logistic or gradient-boosted prediction. Deep learning can extract representations from raw scans; multimodal architectures could fuse cine, Doppler, structured IOTA features and clinical information. For clinically credible evaluation, imaging and clinical predictors must be available before the decision time. Including pathology reports, later referral notes or procedure codes in the predictor set would constitute temporal leakage.

### [Reinforcement learning and scan guidance](#)

Reinforcement learning might eventually support operator-facing suggestions for probe repositioning or selection of additional views, with rewards based on successful standardized visualization. Such a system differs materially from assigning a diagnosis to existing images and would require prospective usability, safety and workload testing. Incorrect guidance during a time-constrained examination can misdirect acquisition and create unrecognized blind spots. Autonomous probe control is therefore not assumed to be ready for routine gynecological care.

### [NLP and large language models](#)

Clinical reports contain text descriptions of lesion size, morphology, recommendations and uncertainty; NLP may transform them into harmonized structured records. Large language models can assist with draft reports and patient-facing explanations, but hallucinated findings and missing negative qualifiers represent safety risks. A language layer should operate only on verified clinician-observed features, disclose AI authorship, retain traceability and require human sign-off. The model should never infer that a lesion is benign because the report does not explicitly mention malignancy.

### [Drug-effect prediction in related diseases](#)

Evidence from predictive treatment models in diabetes, cardiovascular disease, cancer, infections and neurology illustrates general modelling principles—confounding by indication, effect heterogeneity, time-dependent exposure and competing risks—but it is not evidence that gynecological ultrasound AI can predict pharmacologic response in those diseases. Within gynecology, a plausible pathway is to combine baseline lesion measurements with treatment exposure and repeated ultrasound to estimate fibroid response or time to subsequent surgery; this remains a distinct longitudinal causal question from image-based diagnosis.

### Pharmacogenetics and personalized care

Pharmacogenomic or pharmacogenetic traits may eventually help individualize medication regimens, but currently they are not established diagnostic labels for an adnexal ultrasound classifier. Biomarkers such as CA-125 can contribute to malignancy risk assessment in specified settings, although interpretation varies with age, inflammation and benign gynecological disease. Precision medicine here means integrating calibrated imaging, symptoms, laboratory context, reproductive priorities and patient preference—not equating molecular complexity with validated benefit.

### Evaluation of treatment efficacy and effectiveness

An observational comparison of outcomes after AI-assisted ultrasound should clearly specify whether the intervention is image analysis, decision recommendation or a complete pathway including referral. Efficacy can be evaluated under standardized prospective study conditions; effectiveness should measure routine adoption, adherence, downstream testing and outcomes. Discrepancies may arise from clinician override, incomplete interoperability or limited specialist capacity. A treatment-effect model cannot substitute for a randomized impact trial merely because it includes a large number of covariates.

### Causal methods and survival analysis in practice

A target-trial emulation might compare time to definitive diagnosis under AI-assisted versus conventional referral using baseline eligibility and exposure at first ultrasound; propensity-score methods would attempt to balance measurable case mix. Cox models could examine time to diagnosis or intervention and Kaplan–Meier curves visualize follow-up with appropriate censoring. These methods fail when unmeasured severity drives tool use or when only pathology-confirmed patients enter the registry. Time zero, loss-to-follow-up and immortal-time bias must be explicitly addressed.

### Regulatory governance and safety

WHO guidance emphasizes human rights, transparency, accountability and inclusion. FDA guidance now addresses good machine learning practice and use of fit-for-purpose RWE for medical devices. EMA/DARWIN EU is a relevant model for harmonized observational health data, but does not authorize a gynecological diagnostic application. Device-level approvals, intended use, version control, cybersecurity, adverse-event reporting and planned model changes should be reviewed in the jurisdiction of deployment. [10–15]

### Uzbekistan: an actionable implementation programme

For Uzbekistan, a cautious path begins with a Ministry of Health-led inventory of ultrasound equipment, electronic health-record interoperability and pathology follow-

up capacity. Pilot sites should create structured bilingual reporting fields aligned with IOTA, O-RADS and IETA, collect consented de-identified images, document operator training and establish a referral audit. Electronic prescriptions, digital records and registries should be linked only where legally authorized and clinically necessary. Any national system should secure patient data, respect local language usage, and include rural centres and public-sector clinics before claims of nationwide effectiveness.

#### **Uzbekistan: phased prospective evaluation**

A pilot protocol could first run prospective silent-mode external validation; second, a supervised clinician-assistance study; third, a pragmatic randomized or stepped-wedge implementation study after safety review. Baseline metrics should include completeness of structured reports, referral concordance, pathology linkage, false-reassurance rate, time to diagnostic resolution and subgroup outcomes. Clinical pharmacology and pharmacovigilance services can help monitor downstream consequences of decisions, but are not substitutes for device-specific surveillance. Regulatory approval and procurement must be treated as separate decisions.

### **4A. CRITICAL DESIGN SCENARIOS FOR FUTURE RESEARCH**

#### **Prospective multicentre study architecture**

The ideal diagnostic validation cohort consecutively enrolls participants referred for gynecological pelvic ultrasonography across tertiary oncology, general gynecology and community settings. Investigators prospectively register eligibility before knowledge of final outcomes, retain unsuccessful or poor-quality examinations, and define a uniform minimum follow-up for conservatively managed lesions. The index model remains locked throughout validation. Participating centres submit prespecified metadata for scanner manufacturer, probe type, image protocol, sonographer experience and referral indication. Patient-level predictions are time-stamped before access to histology or subsequent clinical decisions. A blinded adjudication committee applies a hierarchy of histopathology, appropriate imaging follow-up and clinical course. This configuration permits estimators that better resemble deployment, rather than a convenient analysis of already operated and diagnostically easy masses. The report should distinguish participants screened, enrolled, imaged, analyzed and verified; exclusions require clinically meaningful reasons.

#### **Benchmarking against expert observation**

When evaluating an algorithm, it is insufficient to compare the AI system with a single untrained assessor or a deliberately reduced clinical risk model. Investigators should prespecify a reference set of expert examiners and standard clinical frameworks, including ADNEX and O-RADS when their intended uses apply. Each observer should

have equivalent access to age, menopausal status, symptoms and clinically relevant ultrasound views. Comparisons may test stand-alone AI, clinician judgment, AI-assisted clinician judgment and structured-rule decision support. Reader randomization and washout may reduce learning or recall effects. The primary outcome should be decision accuracy at a clinically chosen threshold, while net benefit and downstream workload are important secondary outcomes. Discrepant cases ought to be reviewed without disclosing the algorithm output to pathologists. Such designs test a genuine incremental contribution rather than attractive standalone discrimination.

### [Rare malignant and borderline phenotypes](#)

Borderline ovarian tumors, sex-cord stromal tumors, germ cell malignancies and unusual metastatic lesions may be scarce in routine image datasets. An algorithm optimized for abundant benign cysts and common epithelial cancers could exhibit deceptively strong aggregate accuracy while systematically under-recognizing these minority diagnoses. Evaluation must report class counts, confidence intervals, classification thresholds, clinical referral actions and error types. A multicentre consortium can pool eligible uncommon cases, but oversampling in the test set requires calibration and predictive-value analyses reflecting realistic prevalence. Rare histologies also generate uncertainty in annotation, because morphology may overlap with benign subtypes. A cautious reporting system could return indeterminate or specialist-review recommendations rather than assigning an unjustified low-risk probability. Patients may otherwise receive delayed care despite a reassuring score.

### [Uncertainty-aware abstention policies](#)

A clinically reliable algorithm should recognize conditions under which no numerical output is safer than a misleading probability. Such conditions include markedly degraded images, cropped anatomy, missing required clinical inputs, uncommon lesions outside the training distribution and inconsistency between Doppler and grayscale information. Developers can evaluate selective prediction by plotting coverage versus diagnostic error, but abstention itself shifts work toward clinicians. Therefore, the study should report the proportion of patients deferred to expert assessment, their disease prevalence and the time required to complete the pathway. Calibration must be evaluated among accepted and rejected cases separately. A useful output might indicate which mandatory image view is absent and suggest repeat acquisition; a dangerous output would silently discard poor-quality frames and present a highly confident benign classification. Human-readable limits of use are part of device performance.

### Model monitoring and maintenance

Postdeployment surveillance should distinguish data drift, concept drift and workflow drift. Data drift occurs when acquisition settings or the distribution of patient characteristics changes; concept drift occurs when the relationship between findings and diagnoses shifts; workflow drift occurs when staff begin interpreting or acting on model outputs differently. Monitoring should combine image quality, volume of rejected scans, calibration within outcome-verified samples, override frequency and clinically important adverse outcomes. Threshold breaches require a predefined response, such as review, restricted use, recalibration or temporary suspension. Periodic random audit of model-negative patients reduces selective verification. Maintenance must also record software versions and preserve an archive that permits reconstruction of the prediction delivered to each patient. A model updated on recent labels without careful consideration of their selection mechanism can become less safe, not more.

### Clinical impact and patient communication

A gynecological diagnostic support tool should ultimately be judged by whether it supports earlier resolution of diagnostic uncertainty without inducing excessive investigations, unnecessary surgery or inequitable care. Outcomes include time to definitive diagnosis, benign surgeries avoided, missed serious diagnoses, repeat examinations, patient understanding of uncertainty and clinician workload. Patient communication must preserve the distinction between suspicion, calculated risk and confirmed pathology. A patient who hears “the computer ruled out cancer” may interpret a probability far differently from a clinician who understands imperfect sensitivity. Prospective implementation research should include qualitative interviews, informed consent or appropriate notice, incident review and patient-reported experience. A seemingly small change in classification performance can produce large changes in referral burden when deployed across many routine pelvic examinations.

### Privacy, data sharing and cybersecurity

High-resolution ultrasound frames may contain acquisition timestamps, identifiers and machine metadata, and linked reproductive health records are especially sensitive. Data minimization, pseudonymization and encryption are necessary but not sufficient; governance should also describe legal basis for sharing, role-based access, retention, secondary research use and withdrawal limitations. Federated learning may reduce centralization of raw images but still requires secure aggregation, site-level oversight and evaluation of model inversion or membership inference threats. Commercial procurement should specify whether vendors can reuse local images to retrain systems. Security incidents affecting image archives or AI services can interrupt care as well as

breach confidentiality. Thus cybersecurity is a patient-safety issue, not merely an information-technology concern. Regulatory submissions should describe foreseeable misuse and the safeguards for restoring safe conventional workflow.

#### **Fairness and deployment across socioeconomic groups**

A model developed in an oncology referral centre may systematically underperform at remote clinics where image resolution, referral timing and disease prevalence differ. Claims of equitable performance should be supported by stratified results, not only by an inclusive mission statement. A rollout plan could pair community clinics with regional specialist teleconsultation, enabling escalation when model confidence is low. Usability testing must account for language, literacy, disability and variable digital infrastructure. Because smaller subgroups have imprecise estimates, investigators should combine quantitative confidence intervals with qualitative analysis of workflow barriers. Fairness monitoring should continue after deployment: equal average AUC across populations does not guarantee equal rates of missed malignancy at a fixed decision threshold. Local stakeholders must participate in selecting the harms and thresholds considered acceptable.

#### **Reproducibility and model transparency**

A publishable AI diagnostic study should give a versioned data dictionary, eligibility flow, reference standard, image preprocessing pipeline, handling of missing values, train/validation/test separation, hyperparameter search strategy and the precise model checkpoint evaluated. Useful supplementary material includes thresholds, calibration plots, subgroup error matrices and statistical code or sufficiently detailed pseudocode when privacy constraints prevent data release. Proprietary algorithms can still disclose intended use, failure conditions and external validation results. Reporting guidance such as STARD-AI and TRIPOD+AI does not guarantee good performance, but it makes unsupported conclusions easier to detect. Reviewers should reject claims that test cases were fully independent when repeated images from the same participant appeared in model development. Auditable provenance is foundational for scientific trust.

#### **Health-economic evaluation**

Apparent increases in diagnostic sensitivity can improve care but also multiply referrals, biopsies or operations when false-positive rates are not considered. Cost-effectiveness modelling should use local prevalence, clinician time, equipment maintenance, licensing expense, referral transport, pathology capacity and the cost of diagnostic delay. Analyses need plausible ranges rather than invented national unit costs. A conservative pilot may estimate incremental cost per appropriately referred

high-risk patient and compare it with specialist training or standardized ultrasound protocols. Decision-curve analysis complements, but does not replace, economic evaluation because clinical harms and financial costs are distributed differently across patients and facilities. Budget impact over multiple years must incorporate software updates, cybersecurity, staff retraining and postmarket surveillance.

### Sequential diagnostic pathways

Gynecological ultrasound is only one stage of a broader pathway that may include clinical assessment, laboratory markers, repeat scanning, MRI, hysteroscopy, biopsy or surgery. An AI tool inserted at the first stage changes which patients reach later verification and thus can alter the apparent performance of the entire system. Pathway simulation should explicitly represent referral decisions and the probability that each patient completes recommended follow-up. A lesion classified as low risk today may become clinically important if symptoms persist; the protocol should specify revisit criteria. Prospective endpoint adjudication should capture unexpected findings and changes in diagnosis. Evaluation of isolated static images is therefore an incomplete representation of clinical effectiveness.

### Reproductive goals and preference-sensitive decisions

Management of ovarian cysts, fibroids and suspected adenomyosis may depend on fertility intentions, pain burden, pregnancy status and a patient's tolerance for uncertainty. A classifier focused exclusively on malignancy risk could recommend management that is unnecessarily disruptive to reproductive plans if its output is overextended. Decision support should display distinct risk estimates where evidence allows and leave preference-sensitive tradeoffs to clinician-patient discussion. Outcomes of interest can include fertility-preserving management, symptom control, unnecessary invasive procedures and continuity of reproductive care. Researchers must avoid presenting a predictive association as a mandate for treatment. Shared decision-making remains especially important when diagnostic confidence is incomplete and clinical alternatives differ in their consequences.

### Future experimental design and measurable milestones

A translational research agenda should prioritize a locked international multicentre validation of adnexal mass algorithms, followed by prospective randomized studies of decision support where risk-benefit warrants. Endometrial, myometrial and reproductive applications need standardized reference standards before model complexity increases. Study milestones may include structured report completeness, interobserver agreement, calibration, fairness, safe abstention, referral concordance and prospective patient outcomes. An explicit data-quality threshold should be part of

each milestone; unreliable patient linkage or pathology follow-up can invalidate the trial regardless of model sophistication. Ethical approvals, safety committees and clinician training must be planned in advance. Claims of improved effectiveness are warranted only after the relevant endpoint has actually been measured.

## 5. DISCUSSION

### Critical comparison with international evidence

The present synthesis supports a narrow interpretation of technical diagnostic promise. The strongest directly relevant pooled evidence relates to ovarian tumors, but the striking prevalence of high risk of bias limits generalizability. ADNEX models, O-RADS and expert readers remain meaningful baselines. For endometrial and myometrial applications, the smaller and more heterogeneous evidence base argues against extrapolating ovarian AUC values. [1–5]

### Why common study designs exaggerate performance

Case-control enrichment, operated-only cohorts, retrospective convenient sampling and frame-level randomization can produce impressively clean ROC curves. However, actual patients present with ambiguous lesions, competing pathologies and heterogeneous reasons for attending clinics. The model may exploit scanner text overlays, referral centre signatures, data export formats or treatment-related artifacts. Restricting thresholds after examining the test set further overstates accuracy. A clinically responsible paper describes these limitations before making recommendations.

### Bias, data quality and missingness

Bias is not only an abstract fairness metric. Referral bias affects who receives histology; verification bias affects labels; missing history alters prediction; scanner drift changes image distributions; class imbalance masks rare but dangerous errors. Reports should give subgroup denominators, missing-data rates, sample selection flow, adjudication procedures and error review. External evaluation should deliberately stress atypical body habitus, rare tumors, borderline lesions, different reproductive statuses and different image quality levels.

### Explainability and clinical decision support

Explainable AI can improve troubleshooting and human confidence if explanations are stable and clinically valid, but heat maps alone do not validate causal reasoning. Decision-support systems should present probability with confidence and limitations, not an unexplained definitive diagnosis. Human review must remain meaningful: clinicians need authority and sufficient information to reject outputs. Accountability

for patient care should never be obscured by model ownership, procurement agreements or opaque version changes.

### [Ethical AI and regulatory science](#)

The ethical question is not whether AI makes mistakes—it is whether deployment changes the distribution and consequences of mistakes relative to accepted care. WHO principles require attention to autonomy, inclusion and governance; FDA lifecycle frameworks require intended-use-specific evidence and monitoring. Updating a trained model after deployment should trigger change-control review and targeted verification. Governance must specify what data can be reused, who can access outputs and how affected patients can obtain meaningful explanations. [10–15]

### [Limitations of this review](#)

This manuscript is an evidence-informed critical narrative synthesis. It does not claim a reproducible exhaustive search of every subscription database, prospective enrollment, independent meta-analysis or a pooled estimate derived by the present authors. The review includes methodological discussion beyond gynecological ultrasound to respond to broader requirements about RWD, pharmacovigilance and AI, but cross-disease parallels are explicitly conceptual and cannot serve as gynecological effectiveness data. A registered systematic review and a multicentre prospective implementation study are the logical next research stages.

## **5A. TRANSLATIONAL RESEARCH QUESTIONS**

### [External benchmarking in consecutive clinics](#)

A prospective comparison should specify whether the AI model truly improves risk discrimination beyond IOTA expert assessment in consecutive unselected clinics. A common error is to compare published summary AUC values across different patient populations and infer superiority. Rather, participants should undergo both assessments with a uniform reference standard and prespecified decision threshold. A net reclassification improvement statistic is not sufficient if inappropriate referral increases. Clinically meaningful comparison must consider false-negative harm, unnecessary surgery and diagnosis delay. Centre-level effects should be treated explicitly, and variance estimates should not presume independence of scans from the same participant. Transparent trial registration would limit selective reporting of favorable subgroups.

### [Triaging lesions without available histology](#)

When low-risk lesions are not excised, delayed diagnosis and follow-up stability become essential reference-standard components. Short follow-up may label a lesion

benign despite an evolving tumor. Retrospective surgical datasets cannot estimate false reassurance in conservatively managed populations. A robust protocol defines minimum follow-up duration, acceptable imaging changes and escalation criteria before enrollment. A blinded outcome adjudication panel should review ambiguous cases, including loss to follow-up and events outside the original institution. Missing outcome data should be reported and explored using sensitivity analyses. The goal is not to maximize apparent classification performance; it is to determine what happens after the clinician follows the model's recommendation.

### [Clinical interoperability standards](#)

Even a validated image classifier can be clinically ineffective when its outputs do not enter the electronic report, cannot be corrected, or cannot be retrieved during downstream pathology review. The implementation specification should include structured terminology, stable examination identifiers, DICOM-compatible image linkage where supported, interoperable outcome fields and audit logs. Vendor-neutral export and rollback options reduce dependence on one supplier. Each output should retain the model version and the image sequence on which the prediction was made. Clinical users should see whether the input was a single frame or a comprehensive scan, because reliability differs with acquisition completeness. Data interoperability should be evaluated before purchasing systems at scale.

### [Education and sonographer skill](#)

An algorithm may alter how clinicians learn gynecological ultrasound. If trainees only see its final score, they might neglect anatomical reasoning and the acquisition maneuvers that make an examination informative. Educational deployment should therefore present structured descriptors, differential diagnosis and uncertainty rather than simply a colored risk flag. Competency assessments could include unaided scan interpretation, assisted interpretation, acquisition quality and error correction. Educators should compare whether the tool increases knowledge retention or encourages dependence over time. A system that reduces reading time at the cost of weaker independent competence may create long-term fragility in settings where the tool is unavailable or fails.

### [Patient choice and informed consent](#)

Gynecological imaging may be associated with pain, embarrassment, reproductive uncertainty and concerns about cancer. AI-generated explanations can inadvertently increase anxiety if expressed as deterministic disease labels or unsupported percentages. Patient materials should state in clear language what the software analyzes, its limitations, the clinician's responsibilities and how unexpected or

indeterminate findings are managed. Consent for clinical imaging and consent for secondary research reuse are related but distinct. Participation in model development should never be a precondition for standard diagnostic access. Patients should have access to an alternative route when they do not consent to data reuse or algorithmic assistance, consistent with applicable law and local clinical resources.

### Measuring downstream harm

Diagnostic AI research often emphasizes correct labels while omitting downstream harm. For adnexal lesions, false positives may produce anxiety, unnecessary invasive procedures or referral congestion, whereas false negatives may delay diagnosis of malignancy. For leiomyomas and adenomyosis, errors can change decisions about fertility-preserving interventions and symptom management. A patient-safety framework should identify events prospectively and convene independent review for suspected tool-related harm. The comparator is not perfection but the available standard of care, including its own error rate. Researchers should report severity, detectability, reversibility and attribution uncertainty instead of asserting that a harm was definitely caused by the algorithm whenever it followed an incorrect prediction.

### Calibration transfer across facilities

Suppose an AI classifier was calibrated at a specialized referral hospital where malignant adnexal masses are relatively frequent. The same probability may overstate malignancy risk when a community clinic has substantially lower prevalence, even if morphology recognition remains competent. The deployment plan must therefore re-examine calibration and decision thresholds in local prospective cohorts. A recalibration procedure needs enough verified cases, prespecified statistical methods and oversight to avoid overfitting. Transportability cannot be established by recalculating AUC alone. Hierarchical analyses can show how risk relationships vary across centres; they should not conceal sites with clinically important false reassurance.

### Comparative model governance

Hospitals may face claims from multiple vendors that their algorithms have superior accuracy. Procurement committees should request independently audited test cohorts, representativeness of age and menopausal status, subgroup calibration, abstention rates, failure modes and evidence of prospective clinical impact. Comparisons must align intended use and endpoints. A segmentation model for fibroid measurements cannot reasonably be ranked against an ovarian cancer triage model as though they solved the same problem. Pricing should include updates, security audits, staff training, postmarket support and maintenance of a conventional pathway. Governance should

also establish who investigates incidents, how the tool is paused, and whether institutions retain access to audit information after a contract ends.

#### **Longitudinal inference and treatment response**

An imaging model that estimates baseline lesion characteristics might contribute to future models predicting lesion change after hormonal treatment, surgery or expectant management. However, these applications require longitudinal patient records and careful causal design. Patients receiving treatment may differ systematically in symptom severity, fertility intentions and disease burden from untreated patients. Confounding by indication and time-dependent treatment switching can bias naïve associations. For a target trial emulation, eligibility, treatment strategy, time zero, follow-up and outcomes should be defined before analyzing records. Even sophisticated adjustment cannot identify causal effects when clinically decisive confounders were never collected. The present review proposes such designs but reports no original longitudinal effect estimate.

#### **Transparent handling of diagnostic uncertainty**

A medical AI model should communicate more than a point probability. The quality and completeness of acquisition, presence of uncommon findings, stability of relevant inputs and estimated calibration in the patient's clinical setting determine reliability. A risk score without this context may invite unjustified confidence. A transparent interface can explain which data are missing, which lesion was examined and what follow-up action remains indicated regardless of model output. Calibration errors should be treated as a patient-safety problem. Independent clinician interpretation, access to repeat imaging and a documented mechanism for correcting wrong outputs can reduce the harm of uncertainty while maintaining clinical autonomy.

### **6. CONCLUSION**

AI can assist gynecological ultrasound by standardizing lesion descriptions, assisting measurements, prioritizing suspicious adnexal findings and supporting clinical audit. Its clinical value depends on defined intended use, valid reference standards, patient-level multicentre validation, calibrated risk estimates, transparent human oversight and demonstrated net benefit. High retrospective AUC is not a substitute for safety, equity or prospective effectiveness. The evidence is most developed for ovarian mass characterization and less mature for endometrial, myometrial, endometriosis and reproductive imaging tasks. For Uzbekistan, digital infrastructure and pathology-linked reporting should precede ambitious autonomous diagnostic deployment.

## Evidence tables

| Clinical task              | Evidence and comparator                                                                        | Key limitation                           | Practical direction                                |
|----------------------------|------------------------------------------------------------------------------------------------|------------------------------------------|----------------------------------------------------|
| <b>Adnexal masses</b>      | 2025 meta-analysis: ovarian AUC 0.89; ADNEX pooled external AUC 0.93 in low-bias studies [1,3] | 95% studies high risk of bias [1]        | Patient-level external testing; calibration        |
| <b>Endometrial lesions</b> | 2025 endometrial review: 30 studies across different endpoints [5]                             | Heterogeneous reference standards        | Risk-adjusted biopsy pathway trials                |
| <b>Adenomyosis</b>         | Standardized morphology and expert scanning                                                    | Few externally validated studies         | Robust AI Prospective compartment-specific testing |
| <b>Fibroids</b>            | Segmentation and lesion tracking                                                               | Measurement reference error              | Repeatability and management impact                |
| <b>Follicle counting</b>   | Assisted measurement and tracking                                                              | Not a test of natural fertility          | Reader studies and oocyte yield calibration        |
| <b>Device governance</b>   | FDA lifecycle; WHO ethics; STARD-AI [6,10–14]                                                  | Regulatory scope differs by intended use | Postdeployment monitoring                          |

## References

1. Artificial Intelligence Applied to Ultrasound Diagnosis of Pelvic Gynecological Tumors: A Systematic Review and Meta-Analysis. Gynecol Obstet Invest. 2025. Available from: <https://pubmed.ncbi.nlm.nih.gov/40340944/>
2. Timmerman D, Ameye L, Fischerova D, et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. BMJ. 2010;341:c6839. doi:10.1136/bmj.c6839.

3. Barreñada L, Ledger A, Dhiman P, et al. ADNEX risk prediction model for diagnosis of ovarian cancer: systematic review and meta-analysis of external validation studies. *BMJ Med.* 2024;3:e000817. Available from: <https://bmjmedicine.bmj.com/content/3/1/e000817>
4. O-RADS US v2022: An Update from the American College of Radiology's Ovarian-Adnexal Reporting and Data System US Committee. *Radiology.* 2023. doi:10.1148/radiol.230685.
5. Role of artificial intelligence applied to ultrasound in endometrial cancer: a systematic review. *Int J Gynecol Cancer.* 2025. doi:10.1016/j.ijgc.2025.102653.
6. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med.* 2025;31:3283-3289. doi:10.1038/s41591-025-03953-8.
7. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378.
8. Leone FPG, Timmerman D, Bourne T, et al. Terms, definitions and measurements to describe the sonographic features of the endometrium and intrauterine lesions: a consensus opinion from the IETA group. *Ultrasound Obstet Gynecol.* 2010;35:103-112. doi:10.1002/uog.7487.
9. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: WHO; 2021. Available from: <https://www.who.int/publications/b/58847>
10. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. Available from: <https://www.who.int/publications/b/70584>
11. Food and Drug Administration. Good Machine Learning Practice for Medical Device Development: Guiding Principles. Available from: <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/good-machine-learning-practice-medical-device-development-guiding-principles>
12. Food and Drug Administration. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. Final guidance. 2025. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices>

13. European Medicines Agency. Data Analysis and Real World Interrogation Network (DARWIN EU). Available from: <https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/real-world-evidence/data-analysis-real-world-interrogation-network-darwin-eu>
14. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. Available from: <https://www.fda.gov/medical-devices/digital-health-center-excellence/artificial-intelligence-enabled-medical-devices>
15. Food and Drug Administration. Guidances with Digital Health Content. Available from: <https://www.fda.gov/medical-devices/digital-health-center-excellence/guidances-digital-health-content>
16. Food and Drug Administration. CDRH Issues Guiding Principles for Transparency of Machine Learning-Enabled Medical Devices. 2024. Available from: <https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-issues-guiding-principles-transparency-machine-learning-enabled-medical-devices>
17. World Health Organization. WHO releases AI ethics and governance guidance for large multi-modal models. 2024. Available from: <https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models>
18. World Health Organization. Harnessing Artificial Intelligence for Health. Available from: <https://www.who.int/teams/digital-health-and-innovation/harnessing-artificial-intelligence-for-health>
19. World Health Organization. WHO issues first global report on Artificial Intelligence in health and six guiding principles. 2021. Available from: <https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use/>
20. Food and Drug Administration. FDA Use of Real-World Evidence in Regulatory Decision-Making. Available from: <https://www.fda.gov/science-research/real-world-evidence/fda-use-real-world-evidence-regulatory-decision-making>
21. Food and Drug Administration. CDRH and Real-World Evidence. Available from: <https://www.fda.gov/medical-devices/science-and-research-medical-devices/cdrh-and-real-world-evidence>
22. Food and Drug Administration. Town Hall: Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices, Final Guidance. 2026. Available from: <https://www.fda.gov/medical-devices/medical-devices-news-and>

events/town-hall-use-real-world-evidence-support-regulatory-decision-making-  
medical-devices-final-guidance

23. World Health Organization. WHO Consultation towards the development of guidance on ethics and governance of artificial intelligence for health. 2021. Available from: <https://www.who.int/publications/i/item/9789240012752>

24. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance executive summary. 2021. Available from: <https://www.who.int/publications/i/item/9789240037403>

25. EQUATOR Network. STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Available from: <https://www.equator-network.org/reporting-guidelines/the-stard-ai-reporting-guideline-for-diagnostic-accuracy-studies-using-artificial-intelligence/>

26. Food and Drug Administration. Devices Guidances. Available from: <https://www.fda.gov/vaccines-blood-biologics/general-biologics-guidances/devices-guidances>

27. WHO Health Ethics and Governance. Big data and artificial intelligence. Available from: <https://www.who.int/teams/health-ethics-governance/emerging-technologies/big-data-and-artificial-intelligence/>

28. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ*. 2020;370:m3164. doi:10.1136/bmj.m3164.

29. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ*. 2020;370:m3210. doi:10.1136/bmj.m3210.

30. Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and FDA Staff. 2026. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>

Declarations: No patient-level data were collected or analyzed for this review. Funding, conflicts of interest, author contributions and ethics statements must be completed by the actual authors before submission.