

COMPARATIVE DIAGNOSTIC PERFORMANCE OF TRANSVAGINAL ULTRASONOGRAPHY AND MAGNETIC RESONANCE IMAGING IN ADENOMYOSIS: AN EVIDENCE-INTEGRATED REVIEW OF ACCURACY, REAL-WORLD EVIDENCE, AND CLINICAL IMPLEMENTATION

Xolboyeva Mashhura Yoqutjon qizi

Email – yoqutjonovnam@gmail.com

Termez University of Economics and Service

Faculty of Medicine

Abstract

Background: Adenomyosis is a heterogeneous uterine disorder characterized by endometrial-like tissue within the myometrium. Diagnostic uncertainty has consequences for pain management, heavy menstrual bleeding, infertility counselling, and selection of uterus-sparing therapy. Objective: To critically compare transvaginal ultrasonography (TVUS) and magnetic resonance imaging (MRI), and to establish a rigorous framework for evaluating diagnostic performance in routine care. Methods: This integrative evidence review prioritizes comparative diagnostic meta-analyses, the revised Morphological Uterus Sonographic Assessment (MUSA) consensus, clinical guidance, and relevant methodological and regulatory publications. We distinguish published empirical findings from proposed analytical protocols; no original patient cohort, new meta-analysis, or causal treatment-effect analysis was performed. Results: In a 2023 head-to-head meta-analysis encompassing six studies and 595 women undergoing both modalities before hysterectomy histopathology, pooled sensitivity and specificity were 75% and 81% for TVUS versus 69% and 80% for MRI, without statistically significant modality differences ($P = 0.7509$); between-study heterogeneity and selection bias limit universal extrapolation. Revised MUSA terminology separates direct signs, including intramyometrial cysts, echogenic islands, and subendometrial lines or buds, from indirect myometrial responses. MRI improves anatomical mapping when sonography is technically constrained or complex coexisting pathology requires clarification, but junctional-zone thickness thresholds alone are insufficiently robust across physiological contexts. Routine-care datasets may assess diagnostic pathways, interval to treatment, and downstream outcomes, provided verification bias, indication bias, missingness, and treatment changes are addressed. Emerging machine-learning systems remain investigational until externally validated and clinically calibrated. Conclusions: TVUS

should generally be the first-line examination when acceptable and feasible; MRI is a targeted second-line or problem-solving modality. Standardized feature reporting, reference-standard transparency, prospective multicentre validation, and careful governance of imaging artificial intelligence are more consequential than reliance on a single pooled accuracy estimate.

Keywords: adenomyosis; transvaginal ultrasonography; magnetic resonance imaging; MUSA; junctional zone; sensitivity; specificity; diagnostic accuracy; real-world evidence; artificial intelligence; histopathology; clinical decision support; women's health.

1. Introduction

2. Materials and Methods

1.1 Clinical and biological rationale

Adenomyosis encompasses multiple spatial phenotypes rather than a single homogeneous anatomical lesion. Myometrial glandular elements and accompanying smooth-muscle remodelling may be localized or diffuse; disease may coexist with leiomyomas, pelvic endometriosis, prior cesarean scars, and uterine enlargement. These combinations complicate the attribution of heavy menstrual bleeding, dysmenorrhea, chronic pelvic pain, and subfertility to any single finding. Clinical suspicion should therefore remain a hypothesis requiring contextualized imaging rather than a binary diagnosis based on symptoms alone. [1–4]

The historical reliance on hysterectomy histopathology created a particular evidence problem: women receiving a definitive reference-standard diagnosis were disproportionately older, symptomatic, and finished with childbearing. Contemporary fertility-preserving practice necessarily relies on imaging-based probability estimates, where the target condition may be detected without subsequent hysterectomy. The change expands access to diagnosis but also makes the apparent sensitivity and specificity dependent on what investigators call disease and how they verify it. [1–4]

The relevant clinical question is not whether one image modality is intrinsically superior in every setting. It is whether each modality provides the information needed for a specific management decision: recognition of direct myometrial signs, exclusion of alternative structural pathology, description of disease extent, and planning of conservative treatment. A first-line choice should consider availability, patient

preference, contraindications, cost, downstream tests, and the consequences of false reassurance. [2–5]

1.2 Real-world data and evidence as distinct concepts

Real-world data (RWD) are observations generated in the ordinary provision of healthcare and daily life, including electronic health records, administrative claims, disease registries, imaging archives, electronic prescriptions, laboratory databases, wearable-device signals, and patient-reported information. Real-world evidence (RWE) arises only when such data are analysed with a prespecified question, defensible design, and transparent assumptions. A large imaging archive is not itself evidence that MRI improves outcomes; it becomes informative only if eligibility, comparator, reference standard, follow-up, and confounding are specified. [6–10]

For adenomyosis, RWD can document the actual sequence of TVUS, MRI, medical treatment, and surgery, thereby revealing circumstances excluded from narrowly selected accuracy studies. Such datasets can also introduce substantial error: a diagnosis code may represent suspected rather than verified disease; an MRI order may identify clinically complex patients; histology is available principally in those selected for surgery; and absence of a coded symptom cannot be equated with absence of that symptom. These pitfalls make causal identification and data provenance essential. [6–10]

1.3 Randomized trials, efficacy, and effectiveness

Randomized controlled trials (RCTs) reduce confounding for randomized interventions under specific inclusion criteria and protocols. Diagnostic-accuracy research, in contrast, generally compares an index test with a reference standard in a clinically relevant spectrum of patients; routine randomized treatment trials cannot be substituted for paired diagnostic evaluations. Randomization of diagnostic strategies may be useful when the endpoint is patient benefit, such as time to appropriate management, rather than merely the accuracy of an image. [11–13]

Efficacy denotes performance under carefully controlled conditions, whereas effectiveness denotes performance under real-world implementation. A levonorgestrel-releasing intrauterine system, hormonal regimen, or surgical strategy might show efficacy in an RCT but achieve different symptom control in routine practice because of adherence, costs, coexisting fibroids, disease phenotype, and changing reproductive goals. Imaging effectiveness likewise includes operator availability, scan quality, interpretation consistency, referral delays, and whether an imaging result changes care. The distinction is not semantic: it determines the estimand and the interpretation of a clinical study. [7–10]

1.4 Regulation and artificial intelligence

The FDA's RWE program addresses appropriate use of routinely collected data for regulatory decisions, while the European Medicines Agency's DARWIN EU initiative facilitates studies using observational healthcare databases. Their relevance to adenomyosis is methodological, not a claim that either agency has approved an adenomyosis-specific RWD diagnostic algorithm. The World Health Organization's artificial-intelligence guidance emphasizes autonomy, safety, accountability, equity, and protection of health information. [6–10,14]

Artificial intelligence (AI) may help segment the junctional zone, identify MUSA features, and extract imaging findings from narrative reports. However, datasets may contain silent labels that merely reproduce sonographers' opinions; systems trained on hysterectomy-selected cohorts can miscalibrate in younger or fertility-seeking populations. A clinically useful tool needs external validation, subgroup performance reporting, reliable uncertainty quantification, human oversight, and surveillance after deployment. [14–17]

2.1 Review design and transparent source selection

This manuscript is an integrative critical review, not a prospectively registered systematic review. Its empirical anchors are published diagnostic meta-analyses, consensus criteria, and peer-reviewed clinical guidelines. Targeted searches were informed by PubMed-indexed literature, journal publications, and official WHO, FDA and EMA information available through October 2026. Embase, Scopus, Web of Science and the Cochrane Library are recommended sources for a future exhaustive review; independent complete searches of those subscription databases were not undertaken here. Consequently, no fabricated PRISMA flow counts, unpublished pooled estimates, or claims of comprehensive database coverage are supplied.

For a reproducible systematic update, candidate terms include “adenomyosis”, “transvaginal ultrasound”, “sonography”, “magnetic resonance imaging”, “junctional zone”, “MUSA”, “sensitivity”, “specificity”, “diagnostic test accuracy”, and “real world evidence”. Inclusion should prioritize women assessed with both index tests, explicit blinding, a stated and clinically defensible reference standard, extractable two-by-two data, and transparent sampling. Reviews, guidelines and methodological studies are included separately as contextual evidence; case reports and unverified claims should not contribute to pooled estimates. [1–5,11–13]

2.2 Risk of bias, extraction, and synthesis

Data abstraction for a future formal review should capture sample size, care setting, age and reproductive profile, imaging equipment, examiner expertise, MRI sequences, direct and indirect features, reference standard, reference-test interval, fibroid burden, imaging interpretation blinding, and the number of participants not verified by histology. QUADAS-2 is suitable for appraisal of patient selection, index testing, reference testing and flow/timing; applicability should be distinguished from internal bias. STARD supports complete reporting of diagnostic accuracy, while PRISMA-DTA is relevant to a systematic synthesis. [11–13]

Diagnostic performance should be reported with sensitivity, specificity, likelihood ratios, diagnostic odds ratios, confidence intervals and decision-specific predictive values. Sensitivity is the proportion of reference-positive cases recognized by the index test; specificity is the proportion of reference-negative cases correctly identified. Positive and negative predictive values depend on disease prevalence and must not be transferred without adjustment between referral clinics and community populations. A bivariate random-effects model or hierarchical summary ROC model may be appropriate when enough sufficiently comparable studies are identified. [1,4,11–13]

2.3 Causal inference and survival methodology

Propensity-score matching (PSM) balances measured baseline characteristics across groups receiving different diagnostic strategies; it cannot correct unmeasured indication bias. Inverse probability weighting (IPW) forms a pseudo-population in which treatment assignment is less associated with observed baseline covariates, but unstable weights arise under weak positivity. Target-trial emulation specifies a hypothetical clinical trial before analysing RWD: eligibility, treatment strategies, treatment assignment, time zero, follow-up, outcomes, causal contrast, and analysis plan. Failure to align time zero can create immortal-time bias. [18–21]

Kaplan–Meier curves describe observed time-to-event experience subject to censoring assumptions, while Cox proportional-hazards models estimate hazard ratios assuming proportional hazards and suitable covariate adjustment. For adenomyosis, potential endpoints include time from the initial suspected diagnosis to definitive treatment, symptom recurrence, or subsequent hysterectomy; a competing-risk model may be preferable when mutually exclusive reproductive or surgical events occur. These methods are described as tools for future studies, not as calculations conducted in this review. [20–22]

2.4 Artificial-intelligence analysis methods

Machine learning encompasses supervised prediction such as penalized regression, random forests, and gradient-boosted trees; deep learning includes convolutional and transformer architectures that process image textures and spatial patterns. Natural-language processing (NLP) structures free-text sonography and MRI reports; large language models (LLMs) may draft structured summaries but must not be permitted to invent features. Reinforcement learning might theoretically optimize sequential testing decisions, yet its reward specification, safety constraints and deployment evidence remain insufficient for routine adenomyosis care. [14–17]

A valid model development protocol should partition data by patient and preferably by hospital or calendar period, prevent leakage from downstream histology or treatment, prespecify thresholds, report sensitivity and calibration by subgroup, and test robustness across scanners. Explainable-AI heat maps are not evidence of causal reasoning; reader studies and prospective intervention trials are needed to establish improved decisions. [14–17]

3. Results: Critical Synthesis of Published Evidence

3.1 Comparative accuracy: paired evidence versus indirect comparisons

The most directly informative published comparison evaluated women who underwent both TVUS and MRI and then hysterectomy histology. Alcázar and colleagues synthesized six head-to-head studies including 595 women; pooled TVUS sensitivity and specificity were 75% and 81%, with positive and negative likelihood ratios of 3.9 and 0.31. MRI sensitivity and specificity were 69% and 80%, with positive and negative likelihood ratios of 3.5 and 0.39. The reported difference between the modalities was not statistically significant ($P = 0.7509$). These figures pertain to the selected studies rather than all patients with suspected adenomyosis. [1]

The key advantage of a paired design is that disease spectrum is shared between modalities. In an unpaired meta-analysis, an apparently superior MRI estimate might reflect referral patterns and a different prevalence of diffuse disease rather than diagnostic superiority. Even paired studies may favor either test if assessors see the other modality's results, if the interval to pathology is long, or if diagnostic thresholds differ. The underlying meta-analysis identified high patient-selection bias in several studies and substantial heterogeneity. Such limitations justify careful interpretation rather than an assertion of equivalence in every subgroup. [1]

Earlier structured reviews also concluded that TVUS and MRI perform similarly overall, while acknowledging variation by three-dimensional acquisition, operator proficiency and definitions of adenomyosis. Results must be reported by study and design rather than averaged indiscriminately from different publications. [3,4]

3.2 Sonographic MUSA criteria: what constitutes direct evidence

The revised MUSA consensus distinguished direct features reflecting ectopic endometrium within the myometrium from indirect changes that may arise through muscular or architectural responses. Direct features are myometrial cysts, hyperechogenic islands, and echogenic subendometrial lines and buds. Indirect features are globular uterine configuration, asymmetrical myometrial wall thickening, fan-shaped shadowing, translesional vascularity, and an irregular or interrupted junctional zone. [2]

This separation is clinically consequential. A globular uterus can occur with fibroids or other uterine distortion. Fan-shaped shadowing can be influenced by acquisition angle and fibrotic architecture. Conversely, a small direct intramyometrial cyst should not be dismissed merely because global uterine shape appears normal. Doppler imaging may distinguish a cyst from an intramyometrial blood vessel. In the absence of direct features, indirect features alone do not conclusively establish adenomyosis according to the revised MUSA consensus. [2]

Standardized reports should indicate which specific features are present, whether each is direct or indirect, the location and focality of abnormalities, coexisting fibroids, image limitations, and level of reader confidence. Dichotomous report labels without documented features impede later audits and machine-learning research. They also make disagreement difficult to resolve because the reason for the label is invisible. [2,5]

3.3 MRI morphology and junctional-zone interpretation

MRI demonstrates the endometrial–myometrial interface and broader pelvic anatomy using T2-weighted sequences, with selected T1-weighted imaging to detect hemorrhagic foci or characterize associated pelvic pathology. Thickening, irregularity, high-signal foci and distortion of the junctional zone can support adenomyosis, especially in focal posterior-wall or difficult-to-visualize lesions. However, a junctional-zone measurement is sensitive to hormonal state, transient contractions, menstrual phase and scanner technique; isolated numerical thresholds are not an adequate substitute for pattern recognition. [5]

MRI may add value in patients with a large or multifibroid uterus, poor acoustic access, ambiguity between adenomyoma and leiomyoma, or extensive concomitant endometriosis. It can also help the multidisciplinary team map disease prior to complex

uterine-sparing procedures. Yet an additional MRI does not automatically provide independent information if the ultrasound is already high quality and the clinical decision would be unchanged. [1,5]

Reports should note uterine dimensions, lesion location, extent of myometrial involvement, image sequences used, competing diagnoses, and examination limitations. When reporting junctional-zone measurements, radiologists should describe technical and physiological context rather than interpret an isolated cutoff as a universal gold standard. [2,5]

3.4 Dynamic TVUS and anatomical mapping

Dynamic TVUS has a distinctive clinical advantage: the examiner can link focal probe tenderness with uterine anatomy, assess organ mobility and visualize vascular patterns in real time. This cannot prove that a specific lesion is the sole pain generator, but it contributes to a contextual diagnostic assessment. In experienced hands, combined two- and three-dimensional visualization may clarify the coronal plane and junctional-zone continuity. [2,5]

Access to expert sonography remains uneven. A low-cost transvaginal study may be low value if the operator lacks training in direct MUSA features or stores no reviewable cine loops. Conversely, expert TVUS may obviate more expensive cross-sectional imaging for a substantial group with characteristic disease and no management-changing uncertainty. The appropriate comparative outcome is therefore not simply equipment cost but the full diagnostic pathway, including return appointments and unresolved symptoms. [1,2,5]

Patient-centred limitations deserve explicit attention. Transvaginal examination may be declined, contraindicated in a specific situation, distressing, or technically impossible. Transabdominal sonography and MRI then become context-dependent alternatives without implying that a patient must undergo an invasive examination to qualify for investigation. [5]

3.5 Differential diagnosis and mixed pathology

Adenomyosis frequently coexists with leiomyomas and endometriosis. A focal adenomyoma may resemble a fibroid, but diffuse margins, lack of a discrete capsule, translesional rather than circumferential vascularity, and junctional-zone disruption can favor adenomyosis. These are probabilistic features rather than individually definitive signs. Fibroids may distort the uterus and reduce confidence in measuring myometrial thickness or interpreting shadows. [2,5]

Other differential considerations include physiological myometrial contractions, postprocedural changes, scarring and rare malignant pathology when clinical or imaging red flags appear. A suspicious mass, atypical rapid change after menopause, or unexplained abnormal bleeding requires evaluation according to clinical oncology and endometrial-sampling pathways rather than automatic attribution to adenomyosis. Neither TVUS nor MRI alone should be treated as a histologic cancer-exclusion test.

The most useful report identifies coexisting structural abnormalities and explicitly distinguishes confident adenomyotic signs from features potentially explained by another condition. A negative scan cannot always exclude early, mild or focal disease because sensitivity is imperfect and histopathology is not routinely obtained in conservatively managed patients. [1,2]

3.6 The effect of disease spectrum on diagnostic estimates

Diagnostic accuracy is a property of the test operating in a defined population, not an immutable characteristic of the device. Studies of hysterectomy specimens may overrepresent advanced diffuse disease and symptomatic uterine enlargement, making both modalities appear more sensitive than they would be for early or small-volume disease in infertility clinics. At the same time, severe concurrent fibroid disease may reduce the readability of either modality and complicate histological attribution. [1,3,4]

Diagnostic sensitivity can therefore shift with age, parity, hormonal therapy, lesion location and the proficiency of the imaging expert. Subgroup-specific confidence intervals and information about inconclusive scans should be included. Excluding indeterminate examinations from a study may inflate apparent accuracy and conceal the real clinical cost of repeated imaging. [11–13]

Sensitivity and specificity must not be substituted for predictive values. In populations with a low pre-test probability, a positive result for a nonspecific indirect feature may warrant careful review rather than immediate high-intensity therapy. Conversely, a negative result in a patient with strongly suggestive pain and bleeding does not eliminate the need for follow-up and alternative diagnoses. [1,2]

3.7 Clinical decision pathway and resource allocation

A pragmatic diagnostic algorithm begins with clinical history, pregnancy-related considerations when relevant, bleeding assessment, fertility intentions and physical examination as appropriate. TVUS performed using a standardized feature checklist is generally the first imaging choice. When direct MUSA features and the clinical context agree, a working diagnosis may be sufficiently supported for shared management

decisions. MRI is reserved for indeterminate TVUS, complex uterine anatomy, discordant clinical findings, or mapping likely to alter treatment. [2,5]

The value of imaging should be measured by management consequences: whether treatment is changed appropriately, repeat imaging is avoided, preventable procedures are reduced and patient outcomes improve. A modality with slightly higher technical accuracy may be less effective when inaccessible or slow; a cheaper modality may be insufficient when uncertainty causes repeated referrals or inappropriate procedures. Local resource analysis should include reader training, maintenance, scheduling and data interoperability. [6–10]

Fertility-preservation counselling adds another layer. Imaging may characterize extent and concurrent pathology but cannot determine reproductive prognosis with certainty for an individual. Decision-making should balance symptoms, coexisting infertility factors, treatment preference and evolving evidence rather than a single ultrasound or MRI score. [5]

3.8 Real-world data sources for adenomyosis

Electronic medical records can yield presenting symptoms, obstetric history, prior procedures, imaging referrals, pathology results and treatment decisions. Electronic prescription systems provide medication exposure proxies, but a filled prescription does not establish adherence. Laboratory datasets can contribute hemoglobin and anemia assessments in heavy bleeding, while national registries can follow hysterectomy and subsequent healthcare use. Pharmacovigilance reports can describe suspected adverse drug reactions but generally cannot calculate incidence without reliable denominators. [6–10]

Insurance claims are useful for utilization and cost analyses but have imprecise phenotyping and may omit important care outside the insured system. Wearable devices and mobile applications may prospectively capture pain, bleeding, medication reminders and quality-of-life trajectories; missingness may be associated with worse symptoms or disengagement, creating bias. Imaging archives and structured MUSA reports are particularly valuable when images can be independently reread. [6–10]

A linked registry should employ patient-level pseudonymization, standardized diagnostic vocabulary, procedure timestamps, explicit indication for MRI, and documentation of certainty. Data governance must distinguish clinical service provision, research use, and algorithm development permissions. A registry that merely aggregates unverified diagnostic codes cannot resolve disputed imaging accuracy. [6–10]

3.9 RWE and comparative effectiveness of diagnostic strategies

A future target trial could compare “expert TVUS first, MRI only if predefined uncertainty remains” with “routine MRI after TVUS” among newly evaluated patients who meet specified criteria. Time zero must be the initial imaging decision rather than the date of MRI completion. Outcomes might include time to an actionable diagnosis, appropriate change in management, unplanned surgery, patient-reported symptom burden and total direct cost. Important baseline confounders would include disease severity, fibroids, fertility goals, age, facility type and access to imaging. [18–21]

Propensity-score matching and IPW may help handle measured differences, but MRI often follows inconclusive sonography, making confounding by indication particularly severe. Instrumental variables are not automatically valid: facility access to MRI could also affect specialist care, invalidating exclusion assumptions. An observational estimate should present the causal diagram, estimand, positivity diagnostics, weight distributions, balance plots and sensitivity analyses. [18–21]

Kaplan–Meier and Cox methods are appropriate only for correctly defined time-to-event endpoints; they do not estimate ultrasound sensitivity. Competing risks and time-varying interventions may be essential when patients transition between medical and surgical treatment. Results cannot be described as causal merely because a multivariable model returns a statistically significant coefficient. [20–22]

3.10 Machine learning, deep learning, NLP, and LLMs

Machine-learning approaches can integrate MUSA-feature checklists, patient symptoms, hemoglobin levels and ultrasound images to predict the probability of clinically meaningful adenomyosis. Deep-learning models may attempt segmentation of the uterine contours and junctional zone or classification of T2-weighted MRI patterns. Such tasks have distinct ground truths: identifying a reader-labelled feature, reproducing a pathology diagnosis and predicting symptom burden should not be conflated. [14–17]

Natural-language processing can identify direct signs, uncertain language and coexisting fibroids from clinical reports. LLMs can support structured report conversion and patient-facing explanations, but hallucination and loss of negation can produce serious misclassification. Human validation, explicit uncertainty markers and protected clinical data flows are necessary. Reinforcement learning, sometimes proposed for sequential testing, is currently better regarded as a research framework than a clinically validated method for selecting MRI in individual patients. [14–17]

Robust evaluation would require patient-level and site-level separation, external validation on different machines, calibration, subgroup analysis and prospective testing

of whether AI assistance improves expert decisions. A model trained only on MRI studies from tertiary centres may fail on low-resource ultrasound datasets. A high internal AUROC without prospective generalizability does not establish clinical effectiveness. [14–17]

3.11 Treatment forecasting and cross-specialty generalizability

Cross-specialty predictive medicine illustrates possibilities but provides no automatic evidence for adenomyosis treatment response. In diabetes, algorithms may model glucose outcomes; in cardiovascular disease, event risks; in oncology, treatment response; in infectious diseases, resistance; and in neurology, longitudinal disability or relapse. These prediction tasks vary in data-generating process, competing risks and treatment selection, and their performance cannot be transferred to gynecology without independent validation. [14–17]

Within adenomyosis, clinically relevant prediction targets include response to a levonorgestrel-releasing intrauterine system, hormonal therapy discontinuation, recurrent bleeding after conservative intervention, and eventual hysterectomy. Candidate predictors may include baseline symptom severity, age, parity, imaging phenotype, fibroid co-occurrence and treatment adherence. Neither a mechanistic biomarker nor a single MRI measurement currently justifies a universal precision-treatment algorithm. [5]

The methodological priority is defining an actionable decision and evaluating incremental benefit over simpler clinical models. Reporting should quantify calibration, net benefit, fairness and the burden of false treatment predictions, rather than relying exclusively on discrimination metrics. [14–17]

3.12 Safety surveillance, pharmacovigilance, and precision care

Medical management can involve hormonal agents, analgesics and interventions with different safety profiles, so adverse drug reactions (ADRs), safety-signal detection and pharmacovigilance analytics are appropriate adjuncts to a longitudinal adenomyosis registry. Spontaneous reports are valuable for early signal generation but suffer from underreporting, stimulated reporting, duplicate submissions and absent exposure denominators. A signal is a prompt for verification, not a causal effect estimate. [8–10]

Precision medicine would ideally incorporate reproductive goals, bleeding burden, pain phenotype, structural imaging distribution, comorbid conditions and treatment tolerability. Pharmacogenetics or pharmacogenomics could eventually inform drug metabolism or response, but no validated routine pharmacogenomic test currently selects TVUS versus MRI for adenomyosis. Candidate biomarkers must demonstrate reproducibility and clinical utility beyond correlations. [5]

Clinical pharmacology services can help maintain consistent coding of medication exposures, contraindications, adverse events and follow-up. In safety analyses, new-user active-comparator designs and well-defined risk windows are usually more informative than unadjusted comparisons of all treated and untreated patients. [18–21]

3.13 Uzbekistan: actionable digital-health research pathway

A national or multicentre Uzbekistan adenomyosis program should be constructed as a prospective clinical-imaging registry rather than assuming that a complete interoperable national dataset already exists. The Ministry of Health, regional gynecology services, ultrasound departments and radiology specialists could jointly define a minimal data dictionary covering presenting symptoms, ultrasound MUSA signs, MRI indication, imaging outcome, treatment, longitudinal quality-of-life assessment and referral delay. Anonymized imaging archives would enable periodic expert audit.

Existing and planned electronic-health and electronic-prescription initiatives are relevant for linking care episodes and treatment utilization, but operational coverage, data standards and legal authorizations must be checked institution by institution. A phased pilot should first assess data completeness and reader agreement, then diagnostic agreement, and only later comparative strategy effectiveness. Where resources are limited, investments in sonographer training, standardized reports and audit infrastructure may deliver greater early benefit than an indiscriminate expansion of MRI ordering.

Locally developed AI should follow ethical review, cybersecurity evaluation, patient consent or another legitimate legal basis, external validation, monitoring and clearly defined clinical responsibility. Research collaboration across tertiary and district facilities is essential to avoid a model that works only at a well-equipped referral institution. Hospital-specific estimates should not be presented as national disease prevalence. [14–17]

3.14 Heterogeneity, missing data and ethical safeguards

Missing data can be missing completely at random, at random conditional on observed covariates, or not at random; patients with greater symptom burden may have more intense follow-up and disproportionately complete diagnostic labels. Multiple imputation may be reasonable under a defensible missing-at-random model but cannot recover an inaccessible reference standard by assumption. Complete-case analysis should be justified, and sensitivity analyses should test plausible informative missingness. [18–21]

Observer disagreement matters as much as automated classification error. Direct MUSA features can be subtle, and MRI junctional-zone boundaries may vary with acquisition or physiological state. Reader agreement needs prespecified assessment,

while uncertain cases should be retained and reported rather than silently excluded. Structured reporting should capture the complete spectrum of quality and uncertainty. [2,11–13]

Ethical AI requires representativeness, data minimization, transparent performance by demographic subgroup, a mechanism for clinician challenge, and governance against secondary uses of sensitive reproductive information. Explainability should support clinician review but cannot replace demonstrated clinical safety. When the algorithm disagrees with a radiologist or sonographer, the pathway must specify who adjudicates and what additional evidence is required. [14–17]

Table 1. Head-to-head accuracy in histology-verified populations

Index test	Sensitivity	Specificity	LR+	LR–
TVUS	75%	81%	3.9	0.31
MRI	69%	80%	3.5	0.39

Source: Alcázar et al., 2023, six paired studies (n = 595); between-test P = 0.7509. These are published pooled estimates, not calculations performed in this review. [1]

Table 2. Evidence-informed clinical roles

Dimension	TVUS	MRI
Initial examination	Usually preferred first-line	Typically second-line
MUSA feature assessment	Dynamic direct/indirect signs	Complementary morphology and mapping
Complex fibroids	May be technically limited	Can clarify full pelvic anatomy
Patient factors	Transvaginal approach may be declined	MRI access and contraindications matter
Clinical interpretation	Operator expertise critical	Reader expertise and acquisition critical

Interpretive synthesis based on [1,2,5]; this is not a quantitative head-to-head resource-use analysis.

4. Discussion

4.1 What comparison studies prove—and what they cannot prove

The central finding is that the best paired diagnostic synthesis did not identify a statistically significant difference between TVUS and MRI in women selected for

histopathological verification. Absence of a statistically significant difference is not proof of formal equivalence: equivalence requires prespecified margins and a design with adequate precision to exclude clinically important differences. The observed point estimates and confidence around them must be interpreted together with heterogeneity and the differing clinical purposes of each modality. [1]

A policy that reflexively replaces ultrasound with MRI would disregard the imaging expertise that underlies good TVUS performance and the additional logistical costs of cross-sectional imaging. An opposite policy that prohibits MRI even for inconclusive ultrasound could create diagnostic delays in anatomically complex cases. A targeted sequencing strategy reconciles comparable average accuracy with different strengths in diagnostic problem solving. [1,5]

4.2 Bias, verification, and the meaning of a gold standard

Histology remains important for confirmation when tissue is obtained, but hysterectomy sampling imposes a strong selection mechanism. Surgical patients may differ from conservatively managed patients in age, symptom severity, fertility goals and disease extent. This is partial-verification and spectrum bias, not an incidental limitation. Even meticulously conducted accuracy analyses cannot fully extrapolate to early disease if that population is absent from the source studies. [1,11–13]

Imaging-only labels are also imperfect because a second expert opinion may replicate correlated errors rather than deliver an independent reference standard. Better evidence will require longitudinal reference frameworks, expert imaging consensus, surgical pathology when clinically indicated and explicit sensitivity analyses that consider uncertainty in disease status. [2,11–13]

4.3 Explainable AI, regulatory science, and patient safety

WHO guidance requires that clinical AI respect autonomy, promote well-being, operate transparently, foster responsibility, ensure equity and remain sustainable. FDA publications on AI-enabled medical devices emphasize lifecycle risk management and the challenge of models modified after deployment; EMA observational evidence systems create opportunities for postmarket questions but do not replace prospective diagnostic validation. An AI classifier used to prioritize MRI referral would be a clinical decision support tool with potential patient harms if it systematically underestimates disease in underrepresented groups. [8–10,14–17]

Responsible deployment needs version control, predefined performance thresholds, human oversight, monitoring of diagnostic delays, documented escalation for uncertain cases and independent governance of secondary data uses. A saliency map may help audit

where a neural network attends, but it should not be described as a physiological explanation for adenomyosis. Clinical utility must be demonstrated through outcomes, not aesthetic plausibility of highlighted pixels. [14–17]

4.4 Translational and implementation priorities

The next generation of studies should combine representative prospective imaging cohorts, structured MUSA checklists and explicitly defined management endpoints. Separate analyses should examine women seeking fertility treatment, patients with fibroids, and patients whose TVUS is indeterminate. Head-to-head comparisons should include actual nondiagnostic test rates and incremental diagnostic yield after the first test, rather than only sensitivity and specificity among fully readable examinations. [1–5]

In Uzbekistan and similar settings, the most feasible initial improvement is a standardized sonography pathway coupled with access to expert MRI review for selected cases. Before implementing an AI triage system, health services should establish interoperable structured reports, quality review of stored images, informed data governance, and a reliable process for correcting inaccurate labels. These prerequisites are also necessary for credible RWE studies. [6–10,14]

5. Conclusion

TVUS and MRI are complementary, noninvasive modalities with similar aggregate diagnostic performance in available histology-verified head-to-head data. Expert TVUS using revised MUSA terminology is the appropriate first-line strategy for many patients, whereas MRI is most valuable for inconclusive imaging, complex coexisting pathology and treatment-relevant anatomical mapping. The priority for international and Uzbekistan-based research is prospective patient-centred validation, standardized interpretation, rigorous reference-standard reporting and causally defensible analysis of real clinical pathways. Neither RWD volume nor AI model complexity compensates for biased selection, poor labels or missing clinical outcomes.

Declarations: Study type: critical integrative literature review. Original human participants: none. New clinical dataset or statistical meta-analysis: none. Funding: not specified. Conflicts of interest: to be declared by the submitting authors. Data availability: all reported empirical numerical findings are drawn from cited publications.

6. Supplementary Critical Analysis — Pretest probability, likelihood ratios, and decision thresholds

A diagnostic image does not exist outside the clinical context in which it was obtained. A patient with severe progressive dysmenorrhea, heavy menstrual bleeding, a tender enlarged uterus and suggestive previous imaging has a different pretest probability than a person without symptoms in whom a nonspecific myometrial change is noted incidentally. Even when published sensitivity and specificity appear stable, the positive predictive value declines as prevalence falls. A low-probability setting may therefore produce more false-positive labels than a tertiary referral clinic. This is particularly important for indirect MUSA features, which can be mimicked by fibroids or transient myometrial appearances. [1,2]

Likelihood ratios offer a more transportable way to update probability than predictive values, although the underlying ratios must still be appropriate for the target population. A positive likelihood ratio increases the odds of disease by multiplying pretest odds; a negative likelihood ratio reduces them. Because the meta-analytical likelihood ratios reported for both tests are moderate rather than conclusive, neither a positive nor a negative image uniformly settles all clinical decisions. The appropriate next action depends on whether the posterior probability crosses a threshold at which reassurance, additional investigation, symptom-directed treatment, or surgical consultation becomes worthwhile. [1]

Decision thresholds should be explicit in research. A low-risk symptomatic treatment may reasonably proceed without surgical verification, whereas a potentially irreversible intervention demands greater attention to alternative diagnoses and management preferences. An economic model should therefore compare outcomes of entire diagnostic strategies rather than a theoretical maximum of test accuracy. Such a model should avoid using the same study-derived sensitivity, specificity and prevalence as if they were independent, and should propagate uncertainty through probabilistic sensitivity analysis. [6–10]

6. Supplementary Critical Analysis — Disease morphology, menstrual phase, and technical acquisition

Adenomyosis does not have a single macroscopic distribution. Focal disease can create a localized mass-like area; diffuse disease may involve multiple uterine walls; outer-myometrial abnormalities may coexist with pelvic endometriosis. A reporting system that records only “adenomyosis present” erases information relevant to surgical mapping and to the evaluation of infertility and pain. It also obscures the mechanisms underlying

apparent disagreement between MRI and TVUS. When one modality reveals disease extent more clearly, the gain may concern mapping rather than binary disease classification. [2,5]

Image acquisition deserves as much scrutiny as diagnostic definitions. The appearance of the junctional zone changes with uterine contraction and hormonal environment. MRI performed during transient contraction may yield apparent thickening or asymmetry that changes on repeat sequences; ultrasound visualization can vary with insonation angle, probe pressure and the acoustic consequences of fibroids. These effects argue for a structured minimum image set and documentation of quality, not for inflexible use of a millimetre threshold. [2,5]

In comparative studies, investigators should document whether images were read prospectively or retrospectively and whether the radiologist knew the clinical indication or ultrasound findings. Blinding reduces interpretation bias but may be less reflective of actual practice, where clinical context can improve interpretation. Both approaches are valuable if their target question is explicit: blinded accuracy estimates address discrimination from imaging information, while pragmatic evaluations address clinical decision-making with contextual information. [11–13]

6. Supplementary Critical Analysis — Concordant and discordant test results

Clinical concordance is useful but not synonymous with truth. When both TVUS and MRI identify convincing direct signs, confidence may increase; however, their errors are correlated because they visualize related anatomical changes. Multiplying the likelihood ratios of both tests as though their results were statistically independent would overstate confidence. Conversely, disagreement may reflect differing thresholds, lesion visibility, technical constraints or reader expertise rather than an error in one modality. [1–5]

A structured discordance pathway should first re-examine the raw images and the definition of the reported feature. Was a supposed myometrial cyst truly intramyometrial or an obliquely imaged vessel? Was an MRI junctional-zone irregularity caused by a contraction? Did a large fibroid obscure the endometrial–myometrial interface? Expert second review can be preferable to reflexive repeat acquisition where the original images are adequate. If uncertainty persists, follow-up decisions should be guided by symptoms, fertility goals, planned interventions and alternative diagnoses. [2,5]

Future paired investigations can classify discordances by feature type and technical quality and prospectively test whether MRI changed decisions after an inconclusive expert ultrasound. A report that MRI identified more lesions without showing changes in management or patient outcomes would not alone establish a superior clinical pathway.

Similarly, an apparently low discordance rate may conceal exclusion of difficult cases. All indeterminate examinations should be enumerated and analysed to preserve real-world relevance. [1,11–13]

6. Supplementary Critical Analysis — Patient-reported outcomes and longitudinal clinical benefit

The measurement of diagnostic usefulness is incomplete without outcomes that matter to patients. Pain severity, functional interference, bleeding impact, fatigue related to anemia, sexual wellbeing and fertility-related concerns may respond differently to an imaging-informed intervention. Standardized patient-reported outcome measures can complement clinician records, but their timing, recall period and missingness require deliberate design. A single postoperative report may be influenced by the indication for treatment and cannot be attributed solely to the imaging modality used.

A prospective cohort should capture symptom baseline before the index scan and follow participants at clinically meaningful intervals using consistent instruments. Prespecified endpoints could include change in bleeding burden, time to appropriate treatment, need for a second modality, reconsultations and decisional regret. For people seeking pregnancy, pregnancy-related outcomes require separate consideration and adjustment for major fertility determinants. An imaging diagnosis may aid counselling without being a sufficient predictor of reproductive success. [5]

Pragmatic strategy trials could investigate whether expert TVUS with conditional MRI shortens time to meaningful care compared with prevailing referral practice. Randomization should take place at the decision node, not after clinicians have selected a subgroup for MRI based on undocumented reasons. Such trials would answer a different and more clinically useful question than comparisons against hysterectomy pathology alone. RWE studies may complement trials by examining access and adoption in populations that would not meet strict trial criteria. [6–10,18–21]

6. Supplementary Critical Analysis — Data linkage, vocabulary, and reproducibility

Interoperability is not achieved merely by storing images and text in the same electronic folder. An analysis-ready dataset needs stable patient identifiers protected by pseudonymization, trustworthy examination timestamps, modality identifiers, equipment characteristics and consistent feature labels. Reports should distinguish negative findings from features that were not evaluated and from technically uninterpretable regions. A missing MUSA assessment cannot legitimately be converted into an absent feature when training an AI classifier.

Disease codes can be provisional, especially when used for referral and reimbursement. Linkage to imaging narratives and, where available, pathology or longitudinal specialist review can improve phenotyping. An adjudication protocol should document how contradictory records are handled and should permit revisions without overwriting the original clinical record. Version control is necessary because improvements to diagnostic terminology can alter the apparent incidence of adenomyosis even when the underlying population risk remains unchanged. [2,6–10]

Reproducible RWE research benefits from prespecified protocols, transparent data dictionaries, shared analytical code when legally permissible and explicit definitions of exposure windows and follow-up. Independent replication across institutions is particularly valuable for ultrasound studies because operator expertise is not encoded reliably in conventional administrative datasets. Model-development publications should identify the number of unique patients as well as examinations, ensuring that repeated scans from the same person do not contaminate both training and test sets. [14–17]

6. Supplementary Critical Analysis — Statistical uncertainty and research priorities

The interpretation of comparative accuracy depends on the uncertainty surrounding pooled estimates. Confidence intervals express sampling uncertainty under the chosen synthesis model, while prediction intervals, when estimable, communicate the expected range of effects in a new clinical setting. High between-study heterogeneity suggests that a single summary sensitivity cannot replace examination of different imaging protocols and patient populations. A nonsignificant comparison can arise from imprecision rather than evidence that the technologies are interchangeable under all conditions. [1,3,4]

Future systematic reviews should prospectively register their methods, present complete search strings and deduplication rules, document excluded full-text studies, and assess whether threshold differences or clinical populations explain heterogeneity. Meta-regression should be limited to interpretable prespecified modifiers, especially when the number of studies is small. Selection of a subgroup solely because it yields significance invites overfitting and exaggerated certainty. Where publication-level data are too heterogeneous, a structured narrative synthesis is preferable to false precision. [11–13]

The most consequential research gaps include mild adenomyosis in reproductive-age populations, performance in the presence of fibroids, patient acceptance of TVUS, comparative cost effectiveness, diagnostic uncertainty, and inter-reader agreement using revised MUSA definitions. For MRI, standardized sequence acquisition and assessment of physiological confounders merit priority. For AI, multicentre external validation,

calibration, subgroup fairness and prospective effect on clinical decisions should precede marketing claims of superior diagnostic accuracy. [2,5,14–17]

6. Supplementary Critical Analysis — Cross-disease AI models and limits of pharmacological analogies

Methods that predict drug efficacy from RWD in diabetes, cardiovascular disease, oncology, infection and neurology share broad computational tools with gynecological research but not necessarily transferable predictors or clinical causal structures. A glucose time series, tumor mutation signature and uterus image convey different kinds of biological information. Accordingly, benchmarking a gynecologic imaging model against high-performing classifiers in another specialty says little about whether the resulting score changes an adenomyosis management decision. [14–17]

A patient-specific pathway may ultimately combine imaging phenotype with symptom progression and treatment history to forecast benefit from medication or conservative intervention. However, associations between imaging severity and treatment outcome may be confounded by clinician treatment selection. Patients with the most extensive lesions may be preferentially offered surgery, while those with mild symptoms may receive less intense medical therapy. Causal methods can reduce measured confounding but are not a substitute for careful study design or randomization when feasible. [18–21]

Pharmacovigilance analytics require comparable caution. Increased adverse-event reports following a newly publicized therapy may arise from stimulated reporting rather than elevated biological risk. Linkage of prescriptions to documented exposure windows and outcomes can improve interpretation, yet clinical adjudication and sensitivity analyses remain necessary. Precision treatment claims should therefore be reserved for validated decisions that improve patient outcomes, not merely correlations generated by flexible algorithms. [6–10]

6. Supplementary Critical Analysis — Responsible implementation in regional health systems

Any implementation plan must begin with a realistic mapping of available equipment and personnel. High-quality TVUS depends on specialist training, sufficient examination time, storage of reproducible images and periodic case review. MRI may be concentrated at regional or tertiary sites. A tiered referral network should specify clinical situations that trigger secondary imaging so that access is based on uncertainty and management consequences rather than informal influence or ability to pay.

An Uzbekistan pilot could adopt revised MUSA templates at participating clinics, train readers using anonymized case libraries and measure structured-report completion and

interobserver agreement before comparing clinical outcomes. A second phase could link imaging referrals with clinical symptoms and electronic prescriptions, provided institutional authorization and privacy safeguards are established. A third phase could assess whether referral criteria decrease low-value repeat testing without increasing delayed diagnoses. None of these phases should be represented as an existing completed national study without actual data. [2,5–10]

Quality indicators should avoid creating perverse incentives. A very low MRI referral rate may appear cost-effective but could reflect missed complex disease; a high sonography positivity rate might indicate overdiagnosis rather than excellent detection. Balanced metrics should include indeterminate study frequency, second-reader disagreement, interval to appropriate treatment, patient satisfaction and adverse outcomes. Only sustained monitoring can distinguish a genuinely improved diagnostic system from a shift in coding or referral behavior.

An operational diagnostic-accuracy study should prespecify a consecutive or randomly sampled cohort, ensure both modalities are performed within an appropriate interval, mask independent readers to the competing index test, and record every nondiagnostic examination. If histology is available for a selected subset only, investigators should describe the selection mechanism and consider verification-bias analyses. Imaging adjudication is not a perfect substitute for pathology, but longitudinal composite standards can be clearly specified and evaluated with sensitivity analyses. Confidence intervals for each diagnostic statistic and reader-agreement estimates should accompany point estimates. A protocol that retrospectively excludes the most difficult cases may produce an attractive figure while failing to represent clinical decision-making. [1,11–13]

Prospective economic evaluation should capture direct scan expenses, staffing, transport, appointment delays, repeat investigations and downstream treatment decisions. The primary analysis may compare total costs and patient-centred outcomes over a fixed follow-up window, with subgroup analyses for referral-centre access, presence of coexisting leiomyomas, and reproductive intention. Assumptions about test prevalence and pathway adherence should be varied systematically. Better characterization of disease on MRI may be beneficial in a subset without implying that MRI should replace initial ultrasound for all patients. A transparent implementation study should record reasons for exceptions to the pathway and accept patient refusal of transvaginal imaging without penalizing access to care. [1,2,5]

The resulting evidence hierarchy should distinguish a validated diagnostic feature, a pooled test statistic, an association between imaging and outcome, and a claim that modifying the diagnostic pathway improves health. Each step requires different study

designs and assumptions. Clinical investigators, radiologists, sonographers, epidemiologists and patient representatives should participate in the prespecification of outcomes and meaningful decision thresholds. This collaborative approach reduces the chance that success is defined by algorithmic performance alone while clinically important uncertainty remains unresolved.

References

1. Alcázar JL, Vara J, Usandizaga C, Ajossa S, Pascual MA, Guerriero S. Transvaginal ultrasound versus magnetic resonance imaging for diagnosing adenomyosis: a systematic review and head-to-head meta-analysis. *Int J Gynaecol Obstet.* 2023;161(2):397-405. doi:10.1002/ijgo.14609.
2. Harmsen MJ, Van den Bosch T, de Leeuw RA, et al. Consensus on revised definitions of Morphological Uterus Sonographic Assessment (MUSA) features of adenomyosis: results of modified Delphi procedure. *Ultrasound Obstet Gynecol.* 2022;60(1):118-131. doi:10.1002/uog.24786.
3. Tellum T, Nygaard S, Lieng M. Noninvasive diagnosis of adenomyosis: a structured review and meta-analysis of diagnostic accuracy in imaging. *J Minim Invasive Gynecol.* 2020. PMID:31712162. <https://pubmed.ncbi.nlm.nih.gov/31712162/>.
4. Diagnostic accuracy of transvaginal ultrasound and magnetic resonance imaging for adenomyosis: systematic review and meta-analysis and review of sonographic diagnostic criteria. *J Ultrasound Med.* 2021. PMID:33502767. <https://pubmed.ncbi.nlm.nih.gov/33502767/>.
5. Harada T, Taniguchi F, Guo SW, et al. The Asian Society of Endometriosis and Adenomyosis guidelines for managing adenomyosis. *Reprod Med Biol.* 2023;22(1):e12535. doi:10.1002/rmb2.12535.
6. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. Silver Spring (MD): FDA; 2018. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>.
7. US Food and Drug Administration. Real-World Evidence. Silver Spring (MD): FDA. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>.
8. European Medicines Agency. Data Analysis and Real World Interrogation Network (DARWIN EU). Amsterdam: EMA. <https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu>.

9. European Medicines Agency. Real-world evidence. Amsterdam: EMA. <https://www.ema.europa.eu/en/human-regulatory-overview/post-authorisation/real-world-evidence>.
10. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021. <https://www.who.int/publications/i/item/9789240029200>.
11. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009.
12. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi:10.1136/bmj.h5527.
13. McInnes MDF, Moher D, Thoms BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. *JAMA*. 2018;319(4):388-396. doi:10.1001/jama.2017.19163.
14. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. <https://www.who.int/publications/i/item/9789240084759>.
15. US Food and Drug Administration. Artificial intelligence-enabled medical devices. Silver Spring (MD): FDA. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>.
16. 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.
17. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364-1374. doi:10.1038/s41591-020-1034-x.
18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41.
19. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
20. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254.
21. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020. <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.

22. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-481. doi:10.1080/01621459.1958.10501452.
23. Cox DR. Regression models and life-tables. J R Stat Soc Series B Stat Methodol. 1972;34(2):187-220.
24. Guo SW. The pathogenesis of adenomyosis vis-à-vis endometriosis. J Clin Med. 2020;9(2):485. doi:10.3390/jcm9020485.
25. Habiba M, Benagiano G. Classifying adenomyosis: progress and challenges. Int J Environ Res Public Health. 2021;18(23):12386. doi:10.3390/ijerph182312386.
26. Zhai J, Vannuccini S, Petraglia F, Giudice LC. Adenomyosis: mechanisms and pathogenesis. Semin Reprod Med. 2020;38:129-143.
27. World Health Organization. Global strategy on digital health 2020–2025. Geneva: WHO; 2021. <https://www.who.int/publications/i/item/9789240020924>.
28. International Council for Harmonisation. ICH E9(R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Geneva: ICH; 2019. <https://www.ich.org/page/efficacy-guidelines>.
29. European Medicines Agency. Guideline on registry-based studies. Amsterdam: EMA; 2021. <https://www.ema.europa.eu/en/registry-based-studies-scientific-guideline>.
30. US Food and Drug Administration. Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products. Silver Spring (MD): FDA; 2023. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug>.