

Clinical Significance of Anti-Müllerian Hormone and Antral Follicle Count in Ovarian Reserve Assessment: An Integrative Evidence Review of Reproductive Prediction, Real-World Analytics, and Precision Care

Xolboyeva Mashhura Yoqutjon qizi

Email – yoqutjonovnam@gmail.com

Termez University of Economics and Service

Faculty of Medicine

Abstract

Background: Anti-Müllerian hormone (AMH) and antral follicle count (AFC) are widely used to estimate the quantitative component of ovarian reserve, but their interpretation is frequently distorted when ovarian stimulation response is conflated with natural fertility or live-birth probability. Objective: To critically synthesize the clinical utility, technical limitations and implementation implications of AMH and AFC, including their role in assisted reproductive technology, reproductive aging, polycystic ovary syndrome, fertility preservation and real-world evidence generation. Methods: This integrative evidence review uses publicly verifiable professional guidance and peer-reviewed methodological and clinical publications, with emphasis on the 2020 American Society for Reproductive Medicine committee opinion, the updated European Society of Human Reproduction and Embryology ovarian-stimulation guidance and World Health Organization infertility guidance. Sources were interpreted narratively; no de novo participant-level dataset, pooled meta-analysis or prospectively registered systematic search was conducted. Results: AMH and AFC offer clinically useful and broadly comparable estimates of expected oocyte yield and risk of an unexpectedly low or excessive ovarian response. Neither is a stand-alone measure of oocyte competence; chronological age and contextual clinical factors remain central to counseling about pregnancy and live birth. Discordance between assays and ultrasonography, hormonal contraceptive exposure, PCOS, recent ovarian interventions and laboratory platform variation complicate thresholds. Real-world registries may inform calibration and equity of implementation but require careful handling of confounding, missingness, selection and treatment changes. Artificial intelligence can improve image measurement and risk stratification only when externally validated and clinically governed. Conclusions: Clinical benefit comes from

indication-specific testing, standardized analytical procedures, transparent uncertainty and respect for reproductive autonomy, rather than categorical labels based on a single threshold. The proposed Uzbekistan research framework prioritizes longitudinal registries, ultrasound quality assurance, assay traceability, fertility-treatment linkage and patient-centered outcomes.

Keywords

anti-Müllerian hormone; antral follicle count; ovarian reserve; infertility; ovarian stimulation; IVF; assisted reproductive technology; reproductive aging; diminished ovarian reserve; real-world evidence; causal inference; artificial intelligence; fertility preservation; precision medicine

1. Introduction

Conceptual boundaries

Ovarian reserve denotes the remaining pool of ovarian follicles, whereas reproductive potential depends on the much broader joint biology of follicle number, oocyte competence, ovulation, sperm factors, tubal function, endometrial receptivity, lifestyle, reproductive history and time. A biomarker that correlates with the number of oocytes retrieved under pharmacological stimulation need not accurately predict conception without treatment. This distinction is essential to prevent unnecessary anxiety, expensive interventions and misleading claims about a personalized biological clock. Professional guidance cautions against using ovarian reserve testing as a general fertility test. [1,2]

Why two complementary markers matter

AMH is secreted predominantly by granulosa cells of early developing follicles and measured from blood, whereas AFC counts ultrasound-visible small antral follicles across both ovaries. The two measures reflect overlapping yet nonidentical aspects of ovarian biology. AMH is affected by assay calibration, ovarian pathology and medication exposure; AFC depends on imaging equipment, follicle-size definition, examination technique and observer training. Their clinical value therefore emerges from their measurement context, not from apparently precise isolated numbers. [1,2]

Real-world data and evidence

Real-world data (RWD) are routinely collected clinical or patient-generated observations such as electronic records, laboratory results, ultrasound metadata, prescribing histories, treatment-cycle logs and patient-reported outcomes. Real-

world evidence (RWE) comprises inferences about intervention use, effectiveness or safety derived through analysis of RWD. A large registry is not automatically causal evidence: credible RWE needs clear eligibility criteria, endpoint definitions, measurement timing, data completeness and analytical assumptions. FDA and EMA frameworks describe the role of such evidence in regulatory decisions, but do not imply that every observational association establishes clinical effectiveness. [7–10]

[RCTs versus implementation evidence](#)

Randomized controlled trials (RCTs) estimate intervention effects under allocation procedures that, when properly conducted, minimize confounding. They can test whether AMH-guided gonadotropin dosing reduces ovarian hyperstimulation or changes live-birth outcomes. RWD studies can evaluate how such protocols perform across clinics, payer groups, laboratory platforms and populations under routine practice. Yet observational comparisons of AMH-guided versus conventional treatment are vulnerable to confounding by baseline risk, physician preference, treatment accessibility and treatment changes during follow-up. Both designs answer distinct and complementary questions. [7,8,14]

[Efficacy and effectiveness](#)

Efficacy describes performance under controlled trial conditions; effectiveness denotes benefit under ordinary practice. A dosing strategy may have high efficacy in a specialist trial but lower effectiveness when ultrasound expertise varies, medications are unavailable, laboratory turnaround is slow or follow-up is incomplete. By contrast, a simpler pathway relying on a well-performed AFC may yield comparable real-world utility in settings without standardized AMH testing. The appropriate assessment includes costs, safety, patient burden, equitable access, treatment delays and live birth—not predictive discrimination alone. [1,2,8]

[Regulation and computational medicine](#)

WHO stresses fair access to infertility services and responsible AI governance. FDA RWE guidance and EMA DARWIN EU illustrate how structured health data can support evaluation while demanding data reliability. Machine learning may automate follicle counting and integrate multidimensional risk profiles; deep learning may extract sonographic image features; language models may facilitate structured record abstraction. None of these technologies removes the need for careful target definition, external validation, calibration, equity evaluation and appropriate clinical oversight. [3,7–12]

2. Materials and Methods

Study design. This article is a critical integrative narrative review rather than a registered systematic review or original clinical cohort. Evidence was prioritized from ASRM, ESHRE and WHO practice documents, followed by indexed peer-reviewed research and authoritative FDA/EMA methodological guidance. The sources cited were identified through targeted public-source retrieval and reference chaining; a comprehensive reproducible search of subscription-only Embase, Scopus and Web of Science was not performed. Journals including Fertility and Sterility and Human Reproduction were prioritized when relevant; naming a journal or database does not mean all holdings were searched. No invented study counts, pooled effects, participant datasets or database-search flow chart are reported.

Eligibility and appraisal. Included material addressed biological or analytical validity of AMH/AFC, prediction of ovarian response or clinical outcomes, contextual factors (PCOS, hormonal medication, surgery, age), assisted reproduction, causal study designs, digital health, safety or implementation. Greater evidentiary weight was assigned to guideline synthesis, systematic methods and clinical designs with appropriate comparators and endpoints. Exclusions covered unverified claims, noninterpretable cutoff thresholds, commercially promoted tests without independent validation and data unable to separate ovarian response from live-birth prediction.

Synthesis procedure. We separated construct validity (what the assay measures), technical reproducibility, prediction of oocyte yield, prediction of patient-important outcomes, effects of test-guided treatment decisions and implementation feasibility. Where primary sources reported numerical statistics, those values are explicitly attributed rather than re-pooled. In the absence of a comparable endpoint and reference population, numbers were not averaged. Tables offer interpretive matrices instead of fictitious pooled effect estimates.

Propensity score matching (PSM)

Estimate probability of exposure to a particular testing pathway conditional on pre-exposure covariates, match treated and comparator cases on overlap, then inspect standardized mean differences and loss of unmatched individuals. Do not condition on variables measured after treatment initiation or treat matching as eliminating unmeasured confounding.

Inverse probability weighting (IPW)

Weight observations by the inverse estimated probability of received treatment, and for follow-up by inverse censoring probabilities. Check propensity positivity, extreme weights, stabilizing and truncation choices. In IVF registers, time-varying treatments and selection into embryo transfer demand separate consideration.

Cox regression and survival analysis

Define an origin, event, censoring process and risk set before fitting hazards for time to pregnancy, treatment discontinuation or live birth. Check proportional hazards and competing risks; hazard ratios are not simple probability ratios. Recurrent stimulation cycles or competing cessation decisions complicate independence.

Kaplan–Meier methodology

Estimate event-free survival when right censoring is approximately noninformative. Do not use a naive cumulative live-birth curve if fertility treatment discontinuation, competing pregnancy outcomes or informative dropout bias interpretation.

Target-trial emulation

Specify eligibility, treatment strategies (for instance AMH-guided versus AFC-guided dosing), assignment time, grace period, follow-up, outcome, estimand and analysis plan. Align time zero to avoid immortal-time bias. Approximating randomized assignment requires explicit assumptions about confounding.

Causal inference

Use directed acyclic graphs to decide adjustment sets, distinguish total from direct effects, consider collider bias from selecting IVF patients, and report sensitivity analyses for unmeasured confounding. Observed predictive relationships are not themselves causal effects.

Machine learning (ML)

Use regularized regression, gradient boosting or other models to forecast oocyte yield, high response or cancellation. Split by patient and clinic in temporal validation, avoid cycle leakage, and report calibration, discrimination, clinical utility and subgroup fairness.

Deep learning (DL)

Automated ultrasound segmentation/counting may reduce some observer workload but needs reliable labels, external equipment validation, reproducibility audits, quality flags and escalation pathways for atypical ovaries.

Natural language processing (NLP) and large language models

Convert narrative ultrasound and reproductive histories to structured variables only with supervised error audits. Negation, missing sides, measurement units, historic results and copied notes are common failure modes; generated text cannot replace verified data.

Reinforcement learning

Policy optimization for adaptive dosing is exploratory: reward functions must explicitly penalize ovarian hyperstimulation, excessive intervention and inequity. Offline policy evaluation requires behavior-policy overlap and safeguards; no autonomous clinical dosing is recommended.

3. Results: Structured Synthesis of Clinical and Methodological Evidence

3.1 Biological interpretation of AMH

AMH predominantly reflects the number and activity of growing small follicles rather than a direct microscopic census of primordial follicles. A serum result therefore measures an indirect biological phenotype, not remaining reproductive lifespan. Follicular development, granulosa-cell physiology and population differences complicate the mapping from laboratory concentration to anatomical reserve. Biological interpretation is particularly fragile when a testing platform changes or when serial results from different laboratories are juxtaposed as though interchangeable.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.2 AFC as an imaging-derived phenotype

AFC enumerates ultrasound-visible small follicles, often using a two-to-ten-millimeter range under established protocols. The count is jointly determined by follicle visibility, sonographer technique and the actual follicular population. Measurements benefit from defined cycle timing, high-resolution transvaginal scanning where appropriate and explicit bilateral documentation. In patients in whom transvaginal scanning is unsuitable, alternative approaches have different measurement performance and should not be considered interchangeable without validation.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a

seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.3 Relative prediction of ovarian response

ESHRE's updated ovarian-stimulation guidance recommends either AMH or AFC for predicting low and high stimulation response. This recommendation pertains to response stratification, not direct prediction of pregnancy. Quantitative prediction can guide initial counseling and risk mitigation, but cutoff values must be tied to a specific assay, study population, stimulation protocol and response definition. A model accurately classifying extreme oocyte yields may still fail to improve outcomes if acting on its predictions changes only dose without meaningful patient benefit. [2]

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.4 Pregnancy, live birth and patient counseling

ASRM concludes that ovarian-reserve markers are poor independent predictors of natural reproductive potential and do not justify denying infertility treatment. Chronological age provides major prognostic information about reproductive quality that AMH and AFC cannot replace. A low AMH concentration may coexist with viable

oocytes and a successful pregnancy; conversely, a high level does not guarantee embryo euploidy or implantation. Counseling should communicate conditional probabilities rather than label patients fertile or infertile based on an isolated biomarker. [1]

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.5 Discordance between AMH and AFC

The two measurements can diverge because of biological, technical or timing factors, and apparent disagreement may be exaggerated near arbitrary cutoff thresholds. Discordance requires a review of original reports, laboratory units, test platform, follicle-count protocol and prior ovarian surgery rather than reflexive repeated testing. A repeat measurement is clinically defensible only if it can alter a consequential decision and if reproducibility can be improved by using a consistent assay or experienced scanner.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest

probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.6 Age-specific risk assessment

The same absolute AMH or AFC may have different implications at different ages. Population percentiles are not individualized chances of pregnancy, and they may not transport across ethnic, clinical or laboratory populations. Interpretation should start with the reason for testing, age, menstrual and treatment history, and the patient's reproductive priorities. Age-by-marker interactions should be evaluated before adopting uniform action thresholds for broad reproductive populations.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant

medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.7 AMH, AFC and PCOS

Polycystic ovary syndrome frequently features increased small follicle number and elevated AMH, but high reserve markers do not necessarily indicate higher fertility because anovulation may dominate the clinical picture. Use of AMH in adult PCOS diagnostic algorithms has separate requirements from its use for predicting ovarian response. The stimulation context also includes elevated risk of excessive response, reinforcing the need for protocol-specific interpretation and prevention-oriented counseling.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.8 Hormonal contraception and temporal variability

Hormonal contraception may suppress measurable AMH or AFC in some circumstances. Testing undertaken during exposure should be interpreted cautiously and clinical decisions should consider whether retesting after a suitable interval would change treatment. Menstrual-cycle phase matters less for AMH than for some other hormones, but analytical or within-person fluctuations remain possible. Serial trajectories assembled from different assay generations can create false impressions of accelerated follicle loss.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.9 Endometriosis and ovarian surgery

Endometriomas and their treatment can alter ovarian structure and measured reserve. A postoperative decline in AMH may reflect diminished function or tissue injury, yet a single concentration does not fully capture subsequent fertility options. Decision-making must weigh pain, malignancy concern, accessibility of follicles, surgical risk and preservation of tissue. There is a major difference between predicting biochemical change and demonstrating better fertility outcomes after a particular intervention.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.10 Gonadotoxic exposures and fertility preservation

Before potentially gonadotoxic chemotherapy or other ovarian interventions, AMH and AFC can contribute to contextual counseling but cannot precisely forecast individual fertility recovery. Treatment regimen, dose, age and fertility goals remain relevant. Measurement should never be used to delay a time-sensitive cancer treatment without a multidisciplinary plan. Longitudinal post-treatment values require consistent sampling and careful interpretation because temporary treatment-related suppression can be misread as irreversible depletion.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the

result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.11 Assisted reproduction and individual dosing

A patient entering IVF needs a clinically meaningful forecast of low or excessive response, an account of adverse events and transparent explanation of treatment options. Biomarker-guided protocols may rationalize initial dosing, yet larger oocyte numbers are only intermediate endpoints. An ideal treatment decision minimizes ovarian hyperstimulation and unnecessary cancellations while preserving cumulative live-birth opportunity, patient autonomy and affordability. Reporting embryo-transfer-only outcomes can distort the assessment if poorly responding patients are excluded before transfer.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time

would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.12 Biomarker accuracy and data transportability

Predictive discrimination must be interpreted alongside calibration, decision thresholds and local disease prevalence. A model that generalizes within one center may miscalibrate in a population with different age distribution, PCOS burden, laboratory platform or stimulation practice. Every clinical model should be audited in an external setting and periodically recalibrated. Reports should distinguish cross-validated apparent performance from independent prospective evaluation rather than describing both as validation.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.13 Electronic health records and clinic workflow

EHRs can join the indication for AMH testing, menstrual history, ultrasound measurements, hormonal medication and treatment outcome. However, ovarian-reserve testing is often ordered selectively, introducing confounding by indication. Referral-only data may underrepresent women who conceive without treatment. Recording whether values were collected before stimulation, after prior cycles or during contraception is necessary for a scientifically interpretable registry.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.14 Electronic prescribing and pharmacy datasets

Electronic prescriptions allow tracking of gonadotropin type, starting dose, dose escalation, trigger selection and adjunctive therapy. Dispensing is not equivalent to administration or adherence; medication changes require dates. Comparisons of regimens must account for physician decisions informed by AMH and AFC. Treatment-effect questions should not simply compare patients assigned high versus low doses without addressing baseline response risk.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.15 Laboratory databases and assay harmonization

Laboratory systems contribute AMH results, units, assay manufacturer, reference ranges and quality-control flags. Values from different assay generations are not automatically exchangeable. Research-grade harmonization requires explicit version control, calibrated conversion only where scientifically justified and laboratory participation in external quality assurance. Missing platform identifiers can compromise multicenter comparisons even when specimen counts are large.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.16 National reproductive registries

National ART registries can follow initiation, oocyte retrieval, embryo development, transfer and eventual birth, allowing analysis of cumulative outcomes. Registries must capture patients with canceled cycles and disconnected follow-up; otherwise survival analyses inflate success probabilities. Cross-clinic linkage and deduplication matter when patients change providers. Collecting only treatment successes produces severe selection bias.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.17 Insurance data and access inequity

Insurance claims reveal service utilization and financial barriers but do not reliably capture follicle imaging quality or ovarian biology. Reimbursement incentives may alter who is tested, who receives IVF and whose pregnancy outcome remains observed. Socioeconomic measures, referral pathways and regional access therefore belong in models of treatment uptake. Equity audits should examine whether biomarker-based triage disadvantages rural patients or people unable to afford repeated assays.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.18 Wearables, mobile apps and patient-generated evidence

Wearables and reproductive-health apps can contribute menstrual-cycle tracking, treatment symptoms and some longitudinal patient-reported measures. They do not directly count ovarian follicles and should not be marketed as equivalent to AMH or AFC. Self-reported events are vulnerable to selective participation, missingness and device changes. Their greatest research contribution may lie in capturing symptom burden, care continuity and patient priorities between clinic visits.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.19 Pharmacovigilance and ovarian stimulation safety

Safety surveillance should focus on ovarian hyperstimulation syndrome, medication reactions, procedural harms and clinically significant treatment discontinuation. Passive adverse-event reports can detect signals but cannot establish incidence without a denominator. Electronic prescribing and outcomes registries can support stronger active surveillance if exposure, person-time and clinical severity are consistently defined. AMH and AFC serve as contextual risk markers, not pharmacovigilance endpoints in themselves.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the

result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.20 Machine learning for oocyte-yield prediction

Prediction algorithms can integrate age, AMH, AFC, body mass index, stimulation dose, prior response and diagnosis. The design must distinguish baseline models, built from information available before therapy, from dynamic models that incorporate treatment responses. Including actual later-cycle follicle counts in a pretreatment model creates target leakage. Models should report patient-level uncertainty and clinical value, not only a high area under the receiver-operating-characteristic curve.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values

as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.21 Computer vision for follicle counting

Deep-learning systems may assist ovarian ultrasound segmentation and identification of follicles in static frames or cine loops. They are vulnerable to shadowing, overlapping follicles, hardware domain shifts and inconsistent annotation criteria. Expert adjudication, ultrasound phantom quality control and multicenter test sets are necessary. The correct output includes confidence, image-quality flags and recommended manual review where the algorithm cannot count reliably.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.22 Natural language processing and LLMs

Clinic notes often contain AMH, AFC and stimulation outcomes in free text, with duplicate or copied measurements. NLP can retrieve these data but must identify negation, measurement date, laterality and contextual meaning. Large language models may summarize counseling or facilitate registry abstraction, yet hallucinated

quantities or untraceable summaries are unacceptable. Human verification and source-linked extraction should precede any consequential use.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.23 Precision and pharmacogenomics

Pharmacogenetics and genomic variants may eventually refine gonadotropin sensitivity estimates, but clinical utility requires evidence that genotype-informed decisions improve patient outcomes over simpler strategies. A genomic association with ovarian reserve does not validate a medication-dosing rule. Precision reproductive care should combine explanatory biology, clinically calibrated prediction, access constraints and shared decision-making while avoiding premature stratification into deterministic categories.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document

meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.24 Cross-specialty prediction and transfer limits

Methods developed for diabetes, cardiovascular diseases, oncology, infectious disorders and neurology can teach general lessons about robust RWD analysis. Nevertheless, a model trained on drug response in another specialty does not transfer to ovarian-reserve evaluation simply because both involve electronic records. Biological mechanisms, intervention timing, target events, censoring and utility differ. Cross-specialty methods are transferable; disease-specific empirical performance is not.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes

and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.25 Clinical policy scenarios and economic value

Testing policies should compare AMH alone, AFC alone, both tests, and indication-specific no testing against meaningful alternatives. Incremental information may justify dual testing for selected discordant or difficult cases, but routine duplication increases cost and can magnify confusion. Cost-effectiveness must consider local unit prices, specialist availability, procedural safety and loss from unnecessary treatment delays. Without such information, assertions of universal economic superiority are not warranted.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.26 Uzbekistan: proposed interoperable reproductive-health pathway

A feasible implementation roadmap for Uzbekistan would establish a standardized ovarian-reserve minimum dataset, compatible ultrasound reporting templates, assay metadata capture and a consent-based linkage of reproductive services. A pilot could assess between-center differences in AMH platform, AFC reporting and treatment indication before modeling outcomes. Electronic prescriptions and regional referral tracking could enable prospective quality measures. This is a proposal, not a claim that nationwide linked registries currently exist.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.27 Uzbekistan: clinical pharmacology and AI governance

National practice could involve professional societies, the Ministry of Health, laboratory services, reproductive clinics and clinical pharmacology teams in defining minimum data quality and adverse-event monitoring. AI adoption should be phased: retrospective technical validation, prospective silent deployment, clinician-in-the-loop evaluation and post-implementation audits. Data privacy, documentation of model updates, cybersecurity and informed patient communication must be funded alongside software procurement.

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

3.28 Interpretation of recent guidelines

ASRM identifies robust prediction of ovarian response with poor ability to predict unassisted pregnancy. Updated ESHRE guidance endorses AMH or AFC when assessing expected ovarian response. WHO infertility guidance emphasizes evidence-based access to care. These messages are mutually consistent when endpoints remain distinct: a biomarker can be useful for treatment planning and still be inadequate for ruling out fertility. [1–3]

Clinical interpretation: The relevant comparison is not merely whether this information can be measured, but whether acting on the measurement improves outcomes for the specific patient population. Analysis should distinguish pretest probability, measurement error, treatment-selection processes and downstream decisions. A robust protocol should predefine the decision threshold, document meaningful harms from incorrect classification and make uncertainty visible to clinicians and patients. Routine care datasets can test the transportability of the result, provided eligibility, exposure timing and outcome ascertainment are not confused with the biology of reserve. [1,2,7,8]

Research implication: A prospective, multicenter dataset should capture baseline age and reproductive history, the assay platform or ultrasound protocol, relevant medication exposure, initial treatment decision, longitudinal intervention changes and patient-important endpoints. External validation by clinic and calendar time would be required before implementation. The analysis should report subgroup calibration and missing-data patterns, and avoid interpreting post-treatment values as baseline covariates. These methodological safeguards determine whether a seemingly precise risk estimate actually supports safe shared decision-making. [7,9,14]

Domain	Useful endpoint	Important limitation	Practice implication
AMH	Expected oocyte yield / response class	Assay/platform effects; not an oocyte-quality test	Use indication-specific thresholds
AFC	Response prediction and ovarian imaging context	Operator, device and follicle-size variability	Standardized bilateral count
Age	Reproductive counseling and live-birth context	Not adequate as sole stimulation-response marker	Interpret jointly with clinical factors
IVF registry	Cumulative live birth and adverse events	Cancellation and loss-to-follow-up selection	Include all cycle starts
ML imaging	Counting reproducibility	External device and labeling shifts	Human escalation
RWD treatment comparisons	Pragmatic comparative effectiveness	Confounding and missingness	Emulate specified well-target trial

4. Discussion

Synthesis and clinical boundaries

The key paradox is that the best measures of ovarian quantity are often misused as decisive measures of individual fertility. This is neither a laboratory failure nor a

reason to abandon reserve assessment. It is a mismatch between prediction target and clinical claim. AMH and AFC may be appropriate in preparation for stimulation while being inappropriate as general screening tools for the probability of unassisted conception. Reframing the question around planned action prevents the same numeric result from acquiring contradictory meanings.

A practical implication is that measurement protocols and clinical decision pathways must be audited together. If one center uses a different laboratory platform, treatment threshold or counting convention, a difference in apparent ovarian reserve may be operational rather than biological. Model deployment should therefore incorporate calibration monitoring, prospective evaluation, explicit rules for atypical cases and a mechanism for reviewing unintended effects. Reproductive counseling must prioritize patient goals and protect against categorical or stigmatizing language.

[Bias, data quality and missingness](#)

Routine datasets are particularly vulnerable to selection into fertility care: who presents for assessment, who can afford treatment, who proceeds to stimulation, who cancels and who remains observable at birth. Missingness is often informative rather than random. Complete-case analyses may systematically discard unsuccessful or financially constrained pathways. Multiple imputation should reflect a plausible missingness model; sensitivity analysis for nonrandom missingness is needed when key reproductive outcomes may be absent for reasons linked to prognosis.

A practical implication is that measurement protocols and clinical decision pathways must be audited together. If one center uses a different laboratory platform, treatment threshold or counting convention, a difference in apparent ovarian reserve may be operational rather than biological. Model deployment should therefore incorporate calibration monitoring, prospective evaluation, explicit rules for atypical cases and a mechanism for reviewing unintended effects. Reproductive counseling must prioritize patient goals and protect against categorical or stigmatizing language.

[Explainability, ethical AI and clinical decision support](#)

Feature importance is not equivalent to a causal explanation. Explainable AI tools may communicate local drivers of a prediction, but must not obscure calibration failure or create false precision. Decision support should offer the model's intended population, data vintage, threshold and uncertainty, and should permit clinician override with audit logging. Responsible AI principles require clear accountability, safety testing,

data minimization, anti-discrimination safeguards and meaningful patient involvement.

A practical implication is that measurement protocols and clinical decision pathways must be audited together. If one center uses a different laboratory platform, treatment threshold or counting convention, a difference in apparent ovarian reserve may be operational rather than biological. Model deployment should therefore incorporate calibration monitoring, prospective evaluation, explicit rules for atypical cases and a mechanism for reviewing unintended effects. Reproductive counseling must prioritize patient goals and protect against categorical or stigmatizing language.

[FDA, EMA, WHO and regulatory science](#)

FDA promotes principled use of RWD and RWE in product evaluation; EMA has developed a distributed real-world research infrastructure through DARWIN EU. WHO's infertility and AI governance frameworks reinforce equitable access and appropriate oversight. Their roles differ: drug-safety evidence and regulatory RWE do not automatically establish a fertility biomarker's clinical utility. A biomarker pathway should be evaluated on analytical validity, clinical validity and demonstrated net benefit from acting on the test.

A practical implication is that measurement protocols and clinical decision pathways must be audited together. If one center uses a different laboratory platform, treatment threshold or counting convention, a difference in apparent ovarian reserve may be operational rather than biological. Model deployment should therefore incorporate calibration monitoring, prospective evaluation, explicit rules for atypical cases and a mechanism for reviewing unintended effects. Reproductive counseling must prioritize patient goals and protect against categorical or stigmatizing language.

[Limits of this review](#)

This manuscript does not claim exhaustive bibliographic retrieval, protocol registration, participant-level reanalysis or new effect estimates. It is designed as an evidence-based conceptual and implementation synthesis. Consequently, any apparent consensus should be viewed in light of the strength of the underlying guideline and study designs. Future formal reviews should pre-register source strategies and endpoint-specific meta-analysis, while local registry studies should pre-specify causal estimands and handle informative censoring.

A practical implication is that measurement protocols and clinical decision pathways must be audited together. If one center uses a different laboratory platform, treatment threshold or counting convention, a difference in apparent ovarian reserve may be operational rather than biological. Model deployment should therefore incorporate calibration monitoring, prospective evaluation, explicit rules for atypical cases and a mechanism for reviewing unintended effects. Reproductive counseling must prioritize patient goals and protect against categorical or stigmatizing language.

5. Conclusion

AMH and AFC are complementary, clinically useful quantitative markers for estimating ovarian response to stimulation. They are not direct measures of oocyte quality, independent guarantees of conception or valid stand-alone tests of natural fertility. Scientific and clinical progress depends on transparent indications, laboratory and sonographic standardization, endpoint-specific risk communication, appropriate causal designs and longitudinal outcome monitoring. In Uzbekistan, interoperable registries and carefully validated decision support could strengthen reproductive services, provided equitable access and patient rights are preserved. [1–3]

Declarations

Funding: No external funding was declared for this manuscript. Conflicts of interest: To be completed and confirmed by the authors. Ethics: No individual patient data were collected or analyzed for this narrative review. Data availability: Not applicable to this literature synthesis. Author contributions: To be completed by the authors before submission. Review registration: Not registered; not a systematic review.

References

1. Practice Committee of the American Society for Reproductive Medicine. Testing and interpreting measures of ovarian reserve: a committee opinion. Fertil Steril. 2020;114(6):1151-1157. doi:10.1016/j.fertnstert.2020.09.134.
2. European Society of Human Reproduction and Embryology. Ovarian stimulation for IVF/ICSI: guideline update [Internet]. 2025. Available from: <https://www.eshre.eu/OSguideline>

3. World Health Organization. Guideline for the prevention, diagnosis and treatment of infertility. Geneva: WHO; 2025. ISBN 9789240115774. Available from: <https://www.who.int/publications/i/item/9789240115774>
4. International PCOS Network. International evidence-based guideline for the assessment and management of polycystic ovary syndrome [Internet]. 2023. Available from: <https://www.monash.edu/medicine/mchri/pcos/guideline>
5. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021. Available from: <https://www.who.int/publications/i/item/9789240029200>
6. World Health Organization. Regulatory considerations on artificial intelligence for health. Geneva: WHO; 2023. Available from: <https://www.who.int/publications/i/item/9789240078871>
7. US Food and Drug Administration. Real-world evidence [Internet]. Available from: <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>
8. US Food and Drug Administration. Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products. Guidance for Industry. 2023. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products>
9. European Medicines Agency. Real-world evidence [Internet]. Available from: <https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/real-world-evidence>
10. European Medicines Agency. Data Analysis and Real World Interrogation Network (DARWIN EU) [Internet]. 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>
11. European Medicines Agency. Real-world evidence framework to support EU regulatory decision-making: third report. 2025. Available from: <https://www.ema.europa.eu/en/documents/report/real-world-evidence->

framework-support-eu-regulatory-decision-making-3rd-report-experience-gained-regulator-led-studies-february-2024-february-2025_en.pdf

12. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018. Available from: <https://www.fda.gov/media/120060/download>

13. US Food and Drug Administration. Advancing Real-World Evidence Program [Internet]. Available from: <https://www.fda.gov/drugs/development-resources/advancing-real-world-evidence-program>

14. International Council for Harmonisation. E9(R1) addendum on estimands and sensitivity analysis in clinical trials [Internet]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

15. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. doi:10.1371/journal.pmed.0040296.

16. 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.

17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63. doi:10.7326/M14-0697.

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. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-481. doi:10.1080/01621459.1958.10501452.

21. Cox DR. Regression models and life-tables. J R Stat Soc Series B. 1972;34(2):187-220.

22. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4:54. doi:10.1038/s41746-021-00423-6.
23. World Health Organization. Global strategy on digital health 2020–2025. Geneva: WHO; 2021. Available from: <https://www.who.int/publications/i/item/9789240020924>
24. American Society for Reproductive Medicine. Practice guidance [Internet]. Available from: <https://www.asrm.org/practice-guidance/>
25. European Society of Human Reproduction and Embryology. ESHRE Guidelines [Internet]. Available from: <https://www.eshre.eu/Guidelines-and-Legal/Guidelines>
26. National Library of Medicine. PubMed database [Internet]. Available from: <https://pubmed.ncbi.nlm.nih.gov/>
27. Cochrane. Cochrane Library [Internet]. Available from: <https://www.cochranelibrary.com/>
28. US Food and Drug Administration. CDRH and Real-World Evidence [Internet]. Available from: <https://www.fda.gov/medical-devices/science-and-research-medical-devices/cdrh-and-real-world-evidence>
29. European Medicines Agency. Real-world evidence framework to support EU regulatory decision-making: report 2021–2023. 2023. Available from: https://www.ema.europa.eu/en/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained-regulator-led-studies-september-2021-february-2023_en.pdf
30. World Health Organization. Infertility [Internet]. Available from: <https://www.who.int/health-topics/infertility>