Getting Started with AI in Gynaecological Pathology

Path-iQ Getting Started Series · No. 9 of 20 · A practical primer for the practising pathologist

July 2026

1. A subspecialty split between screening and molecular classification

Gynaecological pathology contains two AI opportunities with almost nothing in common. The first is cervical screening, the oldest and most heavily validated application in the entire discipline. The second is endometrial and ovarian carcinoma classification, where the field is being reorganised around molecular categories and where AI's role is to infer those categories from morphology.

Anyone starting here should treat them as separate projects.

2. Cervical: the strongest evidence base in pathology

Cervical cytology screening is covered in detail in the cytopathology primer in this series, but the headline evidence bears repeating because it anchors the whole subspecialty. A system validated across retrospective, prospective and randomised observational cohorts totalling 16,056 participants achieved prospective AUC 0.947, and — critically — cytopathologists assisted by it showed a 13.3% sensitivity improvement over unassisted reading. Population-scale deployment across approximately 700,000 women demonstrated a 5.8% sensitivity gain over manual reading.

For cervical histology, the equivalent task is CIN grading. Distinguishing CIN1 from CIN2 is among the least reproducible calls in pathology, and CIN2 is the threshold at which excisional treatment is typically offered — meaning a reproducibility problem translates directly into over-treatment of young women with fertility implications. p16 immunohistochemistry was introduced precisely to improve this, and automated p16 and Ki-67 dual-stain quantification is a well-suited application: objective, quantitative, and addressing a documented clinical harm.

3. Endometrial carcinoma: morphology inferring molecule

The reclassification of endometrial carcinoma into four molecular groups — *POLE*-ultramutated, mismatch-repair-deficient, p53-abnormal, and no specific molecular profile — has transformed prognostication and is embedded in current WHO classification and treatment guidelines. But molecular classification requires sequencing and immunohistochemistry that are not universally available.

This creates an obvious computational question: can the molecular group be inferred from H&E? Several groups have pursued this, and the task has an appealing structure — four classes, strong outcome separation, and a definitive molecular ground truth for training. Performance to date supports triage rather than replacement: an image-based model can plausibly prioritise which cases most need molecular testing, particularly in resource-constrained settings, but it does not substitute for the assay when treatment depends on it.

Alongside this, grading and histotype assignment in endometrial carcinoma — particularly the endometrioid versus serous distinction in high-grade tumours — remains variably reproducible and is a candidate for algorithmic second opinion.

4. Ovarian and other tasks

Ovarian carcinoma histotyping. The five major histotypes have different biology, chemosensitivity and outcomes, and the high-grade serous versus endometrioid versus clear cell distinctions carry real interobserver disagreement. Automated histotyping has been demonstrated in research cohorts and is one of the better-defined ovarian applications.

Homologous recombination deficiency inference. HRD status determines PARP inhibitor eligibility. Predicting it from morphology is an active research question with the same triage-not-replacement caveat.

Chemotherapy response scoring. Chemotherapy response score in interval debulking specimens is quantitative and variably applied.

Placental pathology. Covered further in the paediatric and perinatal primer, but relevant here: automated assessment of villous maturation, infarction and inflammatory lesions addresses a high-volume, under-resourced specimen type.

Trophoblastic disease. Distinguishing complete from partial hydatidiform mole is a classic morphology-plus-ancillary-testing problem where p57 immunohistochemistry and ploidy analysis do the definitive work.

5. Your first 90 days

Weeks 1–3: decide which stream you are in. If your department runs a cervical screening programme, that is where volume, evidence and funding align, and the project is a cytology project. If you are a tertiary gynae-oncology centre, the molecular classification stream is more relevant.

Weeks 4–8: for the histology stream, baseline your reproducibility on the specific decisions that change management. CIN1 versus CIN2. Endometrioid versus serous. Grade 2 versus grade 3. These threshold calls, not overall diagnostic accuracy, are where assistance pays.

Weeks 9–12: audit your molecular testing pathway. Before deploying a model that infers molecular class from morphology, know what proportion of your endometrial cancers currently receive full molecular classification, and what the barrier is. If the barrier is cost or access, an inference model has a role. If everyone is already tested, it does not.

6. Validation specifics

Hormonal and treatment effect is pervasive. Exogenous hormones, progestin therapy for endometrial hyperplasia, tamoxifen effect and intrauterine devices all alter endometrial morphology substantially, and models trained on untreated tissue misinterpret them.

Cyclical variation is unique to this subspecialty. Normal endometrium changes appearance across the menstrual cycle. A model must have seen the full range or it will call normal secretory endometrium abnormal.

Age and menopausal status shift the baseline for essentially every gynaecological tissue. Confirm training population demographics.

Small biopsy versus hysterectomy: endometrial pipelle samples are fragmented, scant and lack architecture. Most published work uses resection material. This is a substantial transportability gap and should be tested explicitly.

7. Failure modes to anticipate

Atypical hyperplasia versus well-differentiated carcinoma. The distinction that determines hysterectomy versus fertility-sparing management, and one of the least reproducible in the discipline. Noisy training labels are guaranteed.

Serous versus endometrioid at high grade. Morphological overlap is genuine; p53 immunohistochemistry resolves many but not all.

Fragmented pipelle specimens. Architecture-dependent models fail on architecture-free material.

Rare tumours. Gestational trophoblastic neoplasia, uterine sarcomas and sex cord-stromal tumours are individually rare and collectively important. Assume no meaningful model coverage.

8. What to watch next

The convergence of cervical screening programmes onto primary HPV testing changes the cytology workload profile substantially — fewer slides, higher abnormality prevalence, and a triage rather than screening question. Algorithms calibrated for unselected screening will need recalibration, and departments should ask vendors directly about this. On the histology side, watch for validated morphology-based molecular classifiers for endometrial carcinoma reaching regulatory submission; the clinical need in lower-resource settings is genuine and the evidence is maturing.

Vendor due-diligence checklist

Use this before any procurement conversation. The first seven questions apply to every AI tool in pathology; the last three are specific to this subspecialty. A vendor unable or unwilling to answer them in writing has told you something important.

Applies to any pathology AI tool

What is the intended use statement, and what is explicitly excluded from it?

What is the regulatory status in my jurisdiction — cleared, CE-marked, research use only?

Can I see performance on an independent external cohort, not only internal validation?

What is the prevalence of the target condition in the validation cohort versus my practice?

What is the failure mode when the model is uncertain — does it abstain, or always answer?

What monitoring does the vendor provide for performance drift after deployment?

Who is liable if an algorithmic output contributes to a diagnostic error?

Specific to this subspecialty

Was the operating point calibrated on an unselected screening population or an HPV-triage population?

How does it handle exogenous hormone effect, progestin therapy and cyclical variation?

Was it validated on fragmented pipelle specimens or on hysterectomy material?

Ask for answers in writing, and keep them. When the tool underperforms in your laboratory — and at some point it will — the gap between what was claimed and what was validated is where the conversation starts.

9. Key references

#

Source

Note

1

AI enables precision diagnosis of cervical cytology grades, *Nature Communications*, 2024

16,056 participants; +13.3% assisted sensitivity

2

AI-assisted cytology in population-based cervical screening

~700,000 women; +5.8% sensitivity

3

WHO Classification of Female Genital Tumours, 5th edition

Molecular classification of endometrial carcinoma

4

Benchmarking foundation models, *Nature Biomedical Engineering*, 2025

Includes gynaecological tasks

Glossary of terms used in this primer

Whole-slide image (WSI) — a digitised glass slide, typically gigapixel in scale, produced by a scanner and viewed on screen rather than under a microscope.

Foundation model — a large model pretrained on vast unlabelled image data to learn general-purpose visual features, then adapted to specific tasks with comparatively little labelled data.

Weakly supervised learning — training where only a slide-level label is available (for example “contains cancer”) rather than region-by-region annotation. Most clinical pathology AI is built this way, because exhaustive annotation is unaffordable.

AUROC — area under the receiver operating characteristic curve, a single figure summarising discrimination across all thresholds. 0.5 is chance, 1.0 is perfect. It says nothing about calibration or about performance at the threshold you would actually use.

Calibration — whether a model's stated probability matches observed frequency. A well-calibrated model that says "80% confident" is right about 80% of the time. Calibration matters more than accuracy when a clinician acts on the number.

Abstention — the model declining to answer when confidence falls below a threshold, rather than returning its best guess. Essential wherever rare entities are in play.

Batch effect — systematic differences arising from scanner, stain lot, laboratory or site rather than from biology. The leading cause of models performing worse elsewhere than where they were built.

Automation bias — the human tendency to defer to a system that is usually correct, and so to stop checking independently. A property of the workflow, not of the algorithm.

Predetermined Change Control Plan (PCCP) — a regulator-agreed specification of changes a manufacturer may make to an authorised device without a new submission. Its existence acknowledges that these systems are not static.

Shadow deployment — running an algorithm in parallel with routine practice, without acting on its output, to characterise its behaviour on your own material before it touches a live workflow.

Path-iQ — Global Pathology AI Review. Prepared for practising pathologists beginning AI adoption.