Getting Started with AI in Genitourinary Pathology

Path-iQ Getting Started Series · No. 8 of 20 · A practical primer for the practising pathologist

July 2026

1. The subspecialty with the deepest regulatory track record

If you want to understand what mature AI deployment looks like in pathology, look at the prostate biopsy. Every major regulatory milestone in the field has happened here first.

Paige Prostate Detect became the first FDA-authorised AI application in pathology, cleared to aid in the primary diagnosis of prostate cancer. The pivotal study design is worth knowing in detail because it set the template for the field: 16 pathologists reviewed 610 prostate needle biopsy whole-slide images unassisted, then after a washout period reviewed them again with AI assistance. Assisted sensitivity improved by approximately 8% and specificity by 0.7%, with the algorithm correctly classifying every whole-slide image on which a diagnosis was corrected. Gains were larger for non-genitourinary specialists (8.5%) than for GU subspecialists (3.9%) — a finding with obvious implications for where such tools deliver most value. Overall, the study reported roughly a 70% reduction in cancer detection errors.

Independent validation has been broadly supportive, with reported sensitivity above 0.94 and specificity above 0.93 across several groups.

The second milestone was prognostic. ArteraAI Prostate, authorised by De Novo in July 2025, became the first FDA-authorised prognostic whole-slide-image AI device, predicting androgen-deprivation-therapy benefit from a biopsy slide combined with clinical data. This created an entirely new regulatory product code and moved AI from “what is this” to “what should we do about it” — a categorical shift.

2. The task map for GU pathology

Prostate cancer detection and screening. The most mature application in all of pathology. Highest value in high-volume practices and where subspecialist coverage is limited.

Gleason grading and grade group assignment. Gleason grading has well-documented interobserver variability, and it drives active surveillance versus treatment decisions. The PANDA challenge established public benchmarks here, and several models have reported performance at or above expert level. Explainability has been a focus — work published in *Nature Communications* in 2025 developed pathologist-like explainable AI for interpretable Gleason grading, addressing the objection that grading models are opaque.

Tumour quantification. Percentage of core involved and linear tumour extent are numeric values that pathologists estimate. Automated measurement is straightforward and directly feeds risk stratification.

Perineural invasion detection. Discrete, focal, easily missed on rapid review — a good triage target.

Bladder cancer. Deep learning has been shown to predict molecular subtype of muscle-invasive bladder cancer directly from conventional histopathology slides. Grading of non-invasive papillary lesions and assessment of lamina propria invasion are both reproducibility problems where assistance would help.

Renal cell carcinoma. Subtyping and nuclear grading; ISUP/WHO grading is a quantitative assessment applied variably.

Testicular tumours. Low volume, high stakes, and a plausible application for rare-entity decision support rather than automation.

3. Real-world deployment evidence

GU is also where prospective deployment studies exist. The Articulate Pro study, reported in *npj Digital Medicine* in 2026, evaluated AI-assisted prostate biopsy reporting in a live diagnostic setting, with algorithm output displayed in an FDA-authorised viewer. Separate work has assessed whether AI-assisted biopsy diagnosis better predicts radical prostatectomy findings — comparing biopsy Gleason score, grade group and tumour extent against the whole-mount specimen. This is the right kind of evidence: not concordance with another pathologist, but concordance with the definitive specimen.

4. Your first 90 days

Weeks 1–2: quantify your subspecialist coverage gap. The Paige data showing larger gains for non-GU pathologists is the key deployment insight. If your prostate biopsies are read by generalists on a rota, that is your business case. If they are read exclusively by two GU subspecialists, the incremental benefit will be smaller and the case rests on efficiency rather than accuracy.

Weeks 3–6: baseline your grading reproducibility. Circulate 40 archived biopsies among your reporting pathologists. Measure grade group agreement, especially at the 3+3 versus 3+4 boundary, which determines active surveillance eligibility.

Weeks 7–12: shadow-deploy on consecutive biopsies. Log every case where AI and pathologist disagreed on presence of cancer, and every case where grade group differed by one or more. Then review those against subsequent prostatectomy or repeat biopsy where available.

5. Validation specifics

Needle biopsy versus TURP versus prostatectomy are different specimen types with different architecture and artefact. Confirm the intended use.

Post-treatment prostate — after radiotherapy, hormonal therapy or focal ablation — shows morphology that models trained on treatment-naive tissue handle poorly. Treatment effect can mimic and can mask carcinoma.

Benign mimics are the specificity challenge: adenosis, atrophy, partial atrophy and basal cell hyperplasia. Sensitivity-tuned tools will flag these, and your workflow must absorb that without generating reflexive immunohistochemistry.

Immunohistochemistry integration is essential in practice. Most equivocal prostate foci are resolved with a basal cell marker and AMACR. An H&E-only tool cannot participate in that step and should not be expected to.

6. Failure modes to anticipate

Small foci. A single atypical gland is where both humans and algorithms are least reliable, and it is precisely where a second opinion is most valuable.

Grade boundary instability. Models, like humans, are least consistent at the 3+4 / 4+3 boundary — where treatment decisions actually change.

Prevalence effects. In screening populations with lower cancer prevalence than the validation cohort, positive predictive value falls.

Prognostic model transportability. A model predicting therapy benefit was trained in a specific trial population under a specific treatment protocol. Applying it outside that population is an extrapolation, not a validated use.

7. What to watch next

The prognostic lane opened by ArteraAI is the direction of travel: models that predict treatment benefit rather than describe morphology. Expect similar authorisations in other tumour types — the breast clearance in May 2026 already extended the same platform. For GU specifically, watch for AI integration into active surveillance protocols, where reproducible grading and quantification would materially reduce the re-biopsy burden patients currently carry.

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

Is performance reported separately for needle biopsy, TURP and prostatectomy specimens?

How does it behave on post-radiotherapy and post-hormonal-therapy tissue?

What is reported performance at the 3+3 versus 3+4 grade boundary specifically?

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.

8. Key references

#	Source	Note
1	Paige Prostate Detect pivotal validation	16 pathologists, 610 WSI; +8% sensitivity; ~70% error reduction
2	ArteraAI Prostate, FDA De Novo DEN240068, July 2025	First authorised prognostic WSI AI device
3	Articulate Pro study, *npj Digital Medicine*, 2026	Real-world AI-assisted prostate reporting
4	Pathologist-like explainable AI for Gleason grading, *Nature Communications*, 2025;16:8959	Interpretability
5	Woerl et al., *European Urology*, 2020;78:256–264	Molecular subtype of bladder cancer from H&E
6	Applications of AI in Prostate Cancer Care, *ASCO Educational Book*	Slide-level AUCs >0.9 across studies

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.