Getting Started with AI in Surgical Pathology

Path-iQ Getting Started Series · No. 1 of 20 · A practical primer for the practising pathologist

July 2026

1. Why surgical pathology is where AI arrived first

Surgical pathology carries the largest slide volume, the widest differential, and the most consequential sign-outs in the laboratory. It is also the discipline where the digital substrate matured earliest. Once a department scans routinely, every H&E becomes a machine-readable object, and the questions AI can be asked stop being theoretical.

The regulatory record now reflects this. As of the FDA's March 2026 data release, the agency's AI-enabled device list contained over 1,500 authorisations, of which pathology accounts for only a small fraction — an independent audit counted 51 pathology-relevant AI/ML authorisations through April 2026, and only seven of those were whole-slide imaging algorithms. Radiology, by comparison, holds more than a thousand. The gap is not a lack of developer interest. It reflects three binding constraints specific to our discipline: slow lab digitisation, absent data standardisation, and the fact that a pure-software pathology algorithm is reviewed as an in-vitro diagnostic, not as software-as-a-medical-device. That is a materially higher evidentiary bar.

Understanding this asymmetry is the most useful orientation available. The technology is not immature; the deployment infrastructure is.

2. The task map: what algorithms actually do

Almost everything marketed as "AI in pathology" reduces to one of six task classes. Learning to classify a vendor claim into the right box tells you immediately what evidence you should demand.

Detection and triage. The algorithm flags slides or regions likely to contain tumour, ranking a worklist so positives surface first. The lowest-risk, highest-yield entry point. Paige Prostate Detect — the first FDA-authorised AI application in pathology — sits here.

Quantification. Counting or measuring what humans count poorly: mitoses, tumour cellularity, percentage positivity, tumour-infiltrating lymphocytes. AI need not be smarter than you, only more consistent — and on counting tasks it reliably is.

Classification and grading. Assigning a diagnostic category or grade. Higher risk, because the output is closer to the final diagnosis, and grading systems embed clinical convention that a model may learn only imperfectly.

Biomarker inference from morphology. Predicting a molecular result — mutation status, receptor expression, microsatellite instability — directly from H&E. Scientifically the most interesting class and clinically the least mature; performance in independent benchmarks remains well below what a confirmatory assay delivers.

Prognostication. Predicting outcome or treatment benefit from the slide. ArteraAI Prostate, authorised via De Novo in July 2025, was the first FDA-authorised prognostic whole-slide-image device, and it opened a regulatory lane that had previously existed only in the literature.

Workflow and text. Slide–block reconciliation, quality control on scanning, dictation, synoptic report structuring. Unglamorous and frequently the fastest return on investment.

3. Foundation models, and why the vocabulary changed

Until roughly 2023, every pathology AI application was trained from scratch on a task-specific labelled dataset. That era is over. Pathology foundation models are large vision encoders trained self-supervised on enormous unlabelled slide archives, producing general-purpose feature representations that downstream tasks can be built on with far fewer labels.

The scale is worth internalising. UNI was trained on over 200 million patches from more than 350,000 slides at Mass General Brigham; Virchow2 on 3.1 million whole-slide images from over 225,000 patients; Prov-GigaPath on 1.3 billion patches from 171,189 slides across the Providence network, covering 31 tissue types. CONCH took a different route, learning from over a million image–caption pairs to link morphology to language.

What this means practically: a research group or vendor can now build a credible classifier for a rare entity with hundreds of annotated cases rather than tens of thousands. It also means performance claims must be read carefully. A benchmark in *Nature Biomedical Engineering* evaluated 19 histopathology foundation models across 13 cohorts, 6,818 patients and 9,528 slides. The vision-language model CONCH performed best overall, with Virchow2 close behind — but the headline numbers are sobering: mean AUROC around 0.77 for morphological tasks, 0.73 for biomarker prediction, and only 0.63 for prognostic tasks. Foundation models are a powerful substrate, not a finished diagnostic.

4. Your first 90 days

Weeks 1–2: audit your digital readiness. Do you scan? What proportion of cases, at what magnification, on which scanner, into which image management system? An algorithm cannot be deployed onto glass. If your department is not scanning at least a defined case type end-to-end, that is the project — not AI.

Weeks 3–4: pick one problem you can measure. The best first use case has three properties: it is high-volume, it has a known error or variability rate, and its outcome is objectively checkable. Prostate biopsy screening, lymph node metastasis detection, and Ki-67 or HER2 quantification all qualify. “Improve diagnostic accuracy” does not; it cannot be measured.

Weeks 5–8: establish your baseline before you see any AI output. This is the step most departments skip and later regret. Measure your current inter-observer agreement, turnaround time, and deferral or second-opinion rate on the chosen task. Without a pre-AI baseline you will never be able to demonstrate benefit, and you will be vulnerable to a vendor’s benchmark rather than your own.

Weeks 9–12: run a shadow deployment. The algorithm reads every case in parallel; nobody acts on its output. You compare, log discordances, and characterise the failure modes on your own tissue, cut by your

own histology lab, stained on your own platform. Only after this should the tool touch a live worklist.

5. Validating in your own laboratory

Regulatory authorisation is not local validation. The College of American Pathologists' framework for validating whole-slide imaging for diagnostic use is the right mental model, and the same logic extends to algorithms.

Validate on your own cases, not the vendor's, across the full spectrum you actually see — the poorly fixed, the over-stained, the decalcified, the referred-in slides cut elsewhere. Use enough cases that your confidence interval means something; a 20-case validation tells you almost nothing about a 2% error rate. Establish a discordance review pathway before go-live.

Then monitor. Model performance drifts when your scanner is serviced, when a stain lot changes, when a new histotechnologist joins, or when case mix shifts. The FDA now routinely grants Predetermined Change Control Plans — PathAI's AISight Dx clearance in June 2025 included one, allowing specified changes to scanners, displays and file formats without a new submission — which is a regulatory acknowledgement that these systems are not static. Your quality system must not treat them as static either.

6. Failure modes to know before you start

Batch effect and site signature. Models learn scanner and staining characteristics as readily as biology. A model trained at one institution can perform substantially worse at another for reasons that have nothing to do with disease. This is the single most common cause of disappointing external validation.

Automation bias. When a tool is usually right, human readers stop looking independently. Assisted performance studies show gains, but they measure the human–AI system, not the human. Design your workflow so the pathologist forms an impression before the AI output is revealed, at least during the introduction period.

Spectrum bias. Algorithms are typically validated on cohorts enriched for the target condition. Real prevalence is lower, so real-world positive predictive value is lower than the published figure — sometimes considerably.

The rare case. Foundation models are trained on what is common. Your medicolegal exposure lives in what is rare. Assume the algorithm has never meaningfully seen the entity that will get you sued.

7. What to watch next

Three developments will shape the next two years. First, multimodal models fusing morphology with genomics and clinical text — the prognostic ceiling in image-only models suggests that is where the remaining signal lives. Second, the shift from single-task devices to pan-tissue screening tools, which changes the validation question from "does it work for prostate" to "does it work for everything you point it at". Third, report-generating systems that draft synoptic output; these arrive through the LIS rather than the image management system, and need governance most departments have not written.

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

Which specimen types and organ systems is this validated for, and which are outside intended use?

What was the scanner, objective magnification and file format of the training data?

Does the authorisation include a Predetermined Change Control Plan, and what changes does it cover?

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	Chen et al., *Nature Medicine* (UNI)	General-purpose pathology foundation model
2	Zimmermann et al. (Virchow2)	3.1M WSI, 225,401 patients
3	Xu et al., *Nature* (Prov-GigaPath)	1.3B patches, 171,189 WSI
4	Benchmarking foundation models as feature extractors, *Nature Biomedical Engineering*, 2025	

19 models, 6,818 patients

5

AI and digital tools in cancer pathology, *Lancet Digital Health*, 2025

Landscape review

6

Innolitics, AI/ML in Digital Pathology: 2026 Snapshot

51 authorisations, 7 WSI algorithms

7

PathAI AISight Dx 510(k), June 2025

Primary diagnosis clearance with PCCP

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.

Shadow deployment — running an algorithm in parallel with routine practice, without acting on its output, to characterise its behaviour on your own material first.

Path-iQ — Global Pathology AI Review.