Getting Started with AI in Renal Pathology

Path-iQ Getting Started Series · No. 6 of 20 · A practical primer for the practising pathologist

July 2026

1. Why the kidney biopsy is an ideal quantification problem

Renal pathology is built on counting. How many glomeruli are globally sclerosed. What percentage of cortex is fibrotic. How much tubular atrophy. What proportion of glomeruli show crescents, segmental sclerosis, or mesangial hypercellularity. Nearly every scoring system in the discipline — the Banff classification for allograft biopsies, the Oxford classification for IgA nephropathy, the ISN/RPS lupus nephritis classification — converts morphology into a semiquantitative score.

Human beings are poor at this. Interstitial fibrosis and tubular atrophy, one of the strongest predictors of renal outcome, is conventionally estimated by eye in 10% increments, and interobserver agreement on that estimate is only moderate. The same applies to Banff chronicity scoring and to Oxford MEST-C components.

This is the profile where algorithmic assessment is most defensible: a repetitive, quantitative task with an established scoring framework and documented human variability. The AI is not being asked to make a diagnosis. It is being asked to count consistently.

2. The task map for renal pathology

Glomerular detection, segmentation and classification. The foundational task. Detect every glomerulus in the section, then classify each as normal, globally sclerosed, segmentally sclerosed, or showing specific lesions. This has been demonstrated repeatedly across research cohorts and underpins essentially everything else.

Interstitial fibrosis and tubular atrophy quantification. Continuous, reproducible measurement of cortical fibrosis replacing eyeball estimation. Arguably the highest-value single application in the subspecialty because of its prognostic weight.

Banff lesion scoring in allograft biopsies. Automated assessment of interstitial inflammation, tubulitis, glomerulitis and peritubular capillaritis. Banff scores drive rejection treatment decisions and are known to be variably applied between centres. The Banff community has been actively engaged with digital pathology standardisation, which matters — algorithmic scoring is only useful if the underlying definition is agreed.

Immunofluorescence quantification. Automated intensity and distribution assessment of immunoglobulin and complement deposits, a task currently reported in a semiquantitative scale that is essentially operator-calibrated.

Electron microscopy measurement. Automated glomerular basement membrane thickness measurement and foot process effacement quantification. Both are tedious, both are measurable, and both are performed inconsistently.

Adequacy assessment. Automated glomerular counting to determine whether a biopsy meets adequacy criteria — a simple, immediately deployable quality-control application.

3. Where to begin

The kidney biopsy is a small specimen with a high stain burden — H&E, PAS, silver, trichrome, plus immunofluorescence and often electron microscopy on the same case. This has two consequences for anyone starting out.

First, your scanning workflow must handle multiple stains per case and keep them associated. Most published renal AI work uses PAS or silver rather than H&E, because the structures being segmented are defined by basement membrane staining. Confirm which stain a given algorithm expects; an H&E-trained glomerular segmenter will underperform on silver.

Second, the small specimen size makes sampling variability a first-order concern that AI cannot fix. A model that counts 8 glomeruli perfectly is still reading an inadequate biopsy.

Start with IFTA quantification or glomerular counting. Both are objective, both have clear ground truth, and both can be validated retrospectively against archived cases with known outcomes.

4. Your first 90 days

Weeks 1–3: audit your own reproducibility. Have three pathologists independently score IFTA and Banff chronicity on 30 archived cases. Almost every department that does this is surprised by the spread. This exercise is the strongest internal argument for algorithmic assistance you will ever produce, and it establishes the baseline against which any tool must be judged.

Weeks 4–8: scan a retrospective allograft or native biopsy cohort with outcome data. Renal pathology has an advantage here — nephrology follow-up data is unusually complete, so you can test whether an automated fibrosis measure predicts eGFR trajectory better than your semiquantitative score.

Weeks 9–12: run the algorithm in shadow against live cases, focusing on whether automated scores would have changed clinical management. In transplant work, a Banff score change across a treatment threshold is the outcome that matters.

5. Validation specifics

Stain protocol variation is the dominant technical confounder, more so than in most subspecialties, because PAS and silver protocols differ substantially between laboratories and directly affect the boundaries the model segments.

Case mix matters. Native and transplant biopsies present different lesion spectra. Medical renal disease in a diabetic population differs from a lupus-heavy referral practice. Validate on the population you serve.

Section thickness affects apparent cellularity and matrix. Renal biopsies are conventionally cut thinner than routine surgical specimens, and a model trained at one thickness may misjudge another.

6. Failure modes to anticipate

Cortex versus medulla. Fibrosis quantification must be restricted to cortex. A model that includes medulla systematically over-reports.

Edge and tangential glomeruli. Partially sectioned glomeruli at tissue edges are frequently miscounted, and in a 10-glomerulus biopsy a two-glomerulus error is a large proportional error.

Rare lesions. Crescents, thrombotic microangiopathy and amyloid are uncommon in training data relative to their clinical importance.

Scoring is not diagnosis. Automated Banff scoring does not produce a rejection diagnosis; the integration of histology, C4d, donor-specific antibody and clinical context remains the pathologist’s. Tools that present a score as a diagnosis should be treated with suspicion.

7. What to watch next

The most significant development is the drive toward machine-readable, quantitative endpoints in nephrology clinical trials. Regulatory precedent from hepatology — where an AI histology tool has been formally qualified for use in MASH trials — establishes a template that renal trial endpoints are likely to follow, and would create the funding and validation pressure that renal AI has so far lacked. Watch also for integration with spatial transcriptomics from large kidney atlas efforts, which is beginning to link morphological lesions to molecular programmes at cellular resolution.

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 stain was the model trained on — PAS, silver, trichrome or H&E?

Does fibrosis quantification restrict measurement to cortex, and how is medulla excluded?

Was it validated on native biopsies, transplant biopsies, or both?

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

Banff Classification working group publications

Digital pathology and scoring standardisation

2

Oxford Classification of IgA Nephropathy (MEST-C)

Semiquantitative scoring framework

3

AI-assisted fibrosis scoring precedent, *JHEP Reports*, 2026

Cross-organ template for fibrosis quantification

4

Benchmarking foundation models, *Nature Biomedical Engineering*, 2025

Includes kidney among evaluated organs

Note: renal-specific algorithm literature is less consolidated than in oncological subspecialties; readers should verify current tool availability and regulatory status locally before procurement.

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.