Getting Started with AI in Forensic Pathology

Path-iQ Getting Started Series · No. 16 of 20 · A practical primer for the practising pathologist

July 2026

1. A discipline where the evidentiary bar is different

Forensic pathology differs from every other subspecialty in this series in one decisive respect: its output may be tested in court. A diagnostic algorithm in surgical pathology is subject to clinical governance. A forensic algorithm is subject to rules of evidence, adversarial cross-examination, and jurisdiction-specific admissibility standards for scientific testimony.

This changes the calculus completely. An opaque model that improves accuracy may still be inadmissible or, worse, admissible and successfully attacked. Explainability is not a preference in forensic pathology; it is a functional requirement.

Anyone starting here should understand that before assessing any tool.

2. Where the evidence currently stands

A systematic review of AI applications in forensic pathology, forensic medicine, forensic odontology and forensic psychiatry, following PRISMA methodology and covering literature to 2025, identified 65 articles of which 18 met inclusion criteria. The reported performance across domains is instructive:

Deep learning in post-mortem neurological forensic analysis achieved accuracy in the range of 70–94%.

Wound analysis systems, particularly gunshot wound classification, reported accuracy of approximately 88–98%.

AI-enhanced diatom testing for drowning diagnosis achieved precision of 0.9 and recall of 0.95.

Microbiome-based analysis reached accuracy up to 90% for individual identification and geographical origin determination.

Two observations follow. First, the evidence base is genuinely small — 18 qualifying studies across an entire discipline. Second, the reported performance ranges are wide, which reflects heterogeneous tasks and small cohorts rather than mature validation.

Forensic AI is at an earlier stage than most of the subspecialties in this series, and should be presented as such.

3. The task map

Post-mortem interval estimation. One of the oldest problems in forensic medicine and one of the least precise. Machine learning applied to biochemical, microbiological and imaging data has produced models that outperform traditional nomograms in research settings.

Wound characterisation. Gunshot entrance versus exit wound classification, sharp versus blunt force injury, and estimation of wounding instrument characteristics. Among the better-performing published applications.

Drowning diagnosis. Automated diatom detection and quantification. A tedious microscopic counting task with a defined endpoint — a natural automation target, and the reported precision and recall figures are among the strongest in the forensic literature.

Thanatomicrobiology. Microbiome succession after death for post-mortem interval estimation and for geographical origin determination.

Forensic anthropology and odontology. Age and sex estimation from skeletal and dental features, both essentially measurement and pattern-recognition tasks with established reference standards.

Post-mortem imaging. Computed tomography and MRI interpretation, where radiological AI transfers more readily than histological AI.

Sudden cardiac death. Overlaps with cardiovascular pathology; structural and conduction system assessment in the young decedent.

4. Your first 90 days

Weeks 1–4: address admissibility before technology. Consult your jurisdiction’s requirements for expert evidence and scientific methodology. In some jurisdictions a novel method must satisfy defined reliability criteria before opinion evidence based on it is admissible. Establish what standard applies to you, and whether an algorithmically assisted opinion would meet it. This step is not optional and it precedes all others.

Weeks 5–8: choose a quantification task, not an interpretive one. Diatom counting, morphometric measurement and image documentation enhancement are quantitative, explainable and defensible. Cause-of-death inference is none of these.

Weeks 9–12: document everything. Forensic practice already demands chain-of-custody rigour. Extend it to algorithmic processing: record the model, version, parameters, input and output for every case. If you cannot reconstruct in court exactly what the software did, do not use it.

5. Validation specifics

Post-mortem tissue is a distinct domain. Autolysis, putrefaction, post-mortem artefact and variable fixation make forensic histology materially different from surgical histology. Models trained on surgical material should be assumed invalid.

Population specificity is a serious issue in identification tasks. Age and ancestry estimation models trained on one reference population perform poorly on others, and in a forensic context that error has consequences well beyond diagnostic inaccuracy.

Case circumstances are inseparable from interpretation. Forensic conclusions integrate scene information, history and autopsy findings. An algorithm operating on histology alone has less information than the

pathologist and cannot be treated as an independent opinion.

Ground truth is frequently unavailable. Unlike clinical pathology, where a resection or follow-up eventually adjudicates, many forensic questions have no independent confirmation. This limits how strongly any validation study can be interpreted.

6. Failure modes to anticipate

Artefact read as injury. Post-mortem changes, resuscitation artefact, animal predation and insect activity are routinely mistaken for ante-mortem injury by models that have not seen them.

Over-precision. A model that outputs a post-mortem interval of 14.3 hours conveys a precision the underlying science does not support. Reporting such a figure in court invites, and deserves, challenge.

Bias amplification. Where training data reflects historical patterns of investigation or charging, a model can encode and reproduce them. In a criminal justice context this is a serious harm, not a technical inconvenience.

Automation bias under time pressure. High-caseload medicolegal systems create precisely the conditions in which algorithmic output is accepted without independent scrutiny.

7. What to watch next

The most important developments will be procedural rather than technical: professional body guidance on the use of AI in medicolegal opinion, and case law establishing how algorithmically assisted expert evidence is treated. Forensic pathologists should follow these more closely than model performance figures. On the technical side, post-mortem imaging AI is likely to reach practice before histology AI does, since it inherits validation from clinical radiology and produces inherently reviewable output.

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

Does the output meet my jurisdiction’s admissibility standard for novel scientific methodology?

Can the processing be fully reconstructed and explained in court — model, version, parameters, inputs?

Was the model validated on post-mortem tissue, or on surgical tissue?

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	The application of AI in forensic pathology: a systematic literature review, 2025	65 articles screened, 18 included; performance ranges across domains
2	Jurisdiction-specific expert evidence standards	Admissibility of novel scientific methodology
3	Professional body guidance on digital and AI-assisted forensic practice	Consult current national guidance

Note: the forensic AI evidence base is small relative to clinical subspecialties. Claims of performance should be interpreted with corresponding caution, and any deployment must address admissibility before accuracy.

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.