Getting Started with AI in Gastrointestinal and Hepatic Pathology

Path-iQ Getting Started Series · No. 7 of 20 · A practical primer for the practising pathologist

July 2026

1. Two very different opportunities under one heading

Gastrointestinal and hepatic pathology contains the highest-volume specimen type in most laboratories — the colonic biopsy — and simultaneously hosts the single most regulatorily advanced AI histology application anywhere in pathology: automated MASH scoring in liver biopsy.

These are different opportunities and they warrant different strategies. GI is a volume and triage problem. Liver is a precision and reproducibility problem tied to drug development.

2. The liver story: the furthest AI has gone in regulated histology

Metabolic dysfunction-associated steatohepatitis is scored on four features — macrovesicular steatosis, lobular inflammation, hepatocellular ballooning and fibrosis — using the MASH Clinical Research Network system accepted by both FDA and EMA for trial enrolment and endpoints. The problem is that human agreement on these features is poor. Published intra-pathologist agreement figures for repeat trial reads run at approximately 0.72 for steatosis, 0.55 for lobular inflammation, 0.70 for ballooning and 0.72 for fibrosis. Studies have found substantial proportions of enrolled trial cohorts failing to meet entry criteria on re-evaluation by a second hepatopathologist.

This variability has repeatedly confounded MASH drug trials, and it created the commercial and regulatory pressure that AI answered.

AIM-NASH (also reported as AIM-MASH) is a machine learning model trained on more than 100,000 annotations from 59 pathologists who assessed over 5,000 liver biopsies across nine large clinical trials. The EMA approved its use for assisting pathologists in MASH scoring in clinical trials in March 2025. In December 2025 the FDA qualified it as a drug development tool — the first AI tool to receive that qualification. Qualification rested on validation showing AIM-NASH-assisted evaluations aligned well with expert consensus and performed comparably to individual pathologists.

The design detail that makes this work is instructive: the deployment workflow constrains when the pathologist may override the algorithm score, deliberately trading individual judgement for standardisation. That is an unusual and consequential choice, and it is only defensible because the endpoint is trial measurement rather than patient diagnosis.

Comparative work has since examined second-harmonic-generation-based AI digital pathology for fibrosis assessment, and pathologist reads assisted by AI have shown significantly improved concordance for ballooning and inflammation with non-inferior performance on steatosis and fibrosis.

3. The GI story: volume, triage and screening

Colorectal polyp classification. Adenoma versus hyperplastic versus sessile serrated lesion. Enormous volume, well-defined categories, and meaningful interobserver variability on serrated lesions in particular.

Dysplasia in Barrett's oesophagus and inflammatory bowel disease. Among the most reproducibility-challenged diagnoses in pathology, with direct surveillance and surgical consequences. A strong candidate for algorithmic second opinion.

Coeliac disease. Villous height to crypt depth ratio and intraepithelial lymphocyte counting are quantitative tasks scored semiquantitatively by Marsh grade — a natural quantification target.

Helicobacter pylori detection. Simple, high-volume, and frequently requiring special stains. Automated detection on H&E could reduce ancillary stain use measurably.

Microsatellite instability and molecular inference. Predicting MSI status directly from colorectal H&E is one of the most-studied morphology-to-molecular tasks in computational pathology. Performance is respectable in research settings but does not replace confirmatory testing.

Neoadjuvant response assessment. Tumour regression grading in rectal and gastro-oesophageal resections is quantitative and variably applied.

4. Your first 90 days

If your interest is liver: determine whether your department participates in MASH trials. If so, engagement with AI scoring is not optional in the medium term — trial sponsors are moving this way. Begin by auditing your own repeat-read agreement on archived MASH biopsies.

If your interest is GI: start with polyp classification or *H. pylori* detection. Both are high-volume, both have objective ground truth available retrospectively, and both offer measurable workflow benefit — reduced special stain ordering is a direct, quantifiable cost saving that funds further work.

Weeks 9–12: in either stream, run shadow deployment and specifically log cases where the algorithm and pathologist crossed a management threshold — dysplastic versus non-dysplastic, F2 versus F3 fibrosis. Threshold crossings, not average agreement, are what determine clinical impact.

5. Validation specifics

Biopsy versus resection matters greatly in GI. Architecture available in a resection is absent in a pinch biopsy, and models trained on one perform poorly on the other.

Stain dependency is critical in liver. Fibrosis assessment depends on trichrome or picrosirius red; steatosis and ballooning are assessed on H&E. A comprehensive scoring tool needs both, correctly paired.

Fragmentation is routine in GI specimens and disrupts architecture-dependent models. Test explicitly on fragmented material.

Treatment effect is pervasive — proton pump inhibitors, biologics, immune checkpoint inhibitors and chemotherapy all alter GI mucosal morphology in ways that resemble disease. Immune-related colitis in particular is under-represented in older training sets.

6. Failure modes to anticipate

Sessile serrated lesions. The category with the worst human agreement and therefore the noisiest labels.

Sampling in liver. A model can score a biopsy perfectly while the biopsy misrepresents the liver. Fibrosis is heterogeneous; a 15mm core is a small sample.

Inflammation attribution. Distinguishing active IBD from infectious colitis from drug effect requires clinical information the algorithm does not have.

Automation bias in trial reads. Where the workflow restricts override, the pathologist's role becomes quality control. That must be understood and staffed accordingly.

7. What to watch next

The MASH precedent is the story to follow. A regulator qualifying an AI histology tool as a drug development tool creates a template that other fibrotic and inflammatory diseases will follow — renal, pulmonary and cardiac fibrosis endpoints are the obvious candidates. On the GI side, watch for pan-tissue screening tools reaching the colonic biopsy workload, where the volume-to-yield ratio makes even modest triage gains economically significant.

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 this validated on biopsies, resections or both, and is performance reported separately?

For liver tools, which stains are required and how are paired stains associated?

How does it handle immune checkpoint inhibitor-related and drug-induced mucosal change?

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

FDA qualification of AIM-NASH as drug development tool, December 2025

First AI drug development tool qualified

2

EMA approval of AIM-NASH for MASH trials, March 2025

>100,000 annotations, 59 pathologists, >5,000 biopsies

3

Clinical validation of AI-based pathology tool for MASH scoring, 2025

AIM-MASH repeatability vs published manual agreement

4

AI-based automation of enrollment criteria and endpoints in liver disease, *Nature Medicine*, 2024

Trial enrolment variability

5

AI-assisted fibrosis scoring in MASH (SHG-based), *JHEP Reports*, 2026

Pathologist decision-making under AI assistance

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.