Getting Started with AI in Hematopathology

Path-iQ Getting Started Series · No. 3 of 20 · A practical primer for the practising pathologist

July 2026

1. Why hematopathology is a different problem

Hematopathology is the most data-integrative discipline in the laboratory. A single diagnosis of acute myeloid leukaemia routinely fuses morphology, immunophenotype by flow cytometry, cytogenetics, and a next-generation sequencing panel — then places the result against a classification system that is itself being rewritten around genomics. The American Society of Hematology's 2025 position paper in *Blood* framed this precisely: the diagnosis and treatment of haematologic disorders depend on integrating imaging, pathology, omics and laboratory parameters, and that integration is where machine learning has the clearest role.

This makes hematopathology unusual. In most subspecialties the AI question is "can a model read this slide". Here it is "can a model reason across five modalities that currently reach the pathologist in five different systems".

2. The task map for haematology

Peripheral blood and bone marrow differential counting. The most mature application by far. Automated digital morphology systems that pre-classify white cells and present them for technologist confirmation have been in routine laboratory use for years, and the hematology analyser category is the single largest contributor to the FDA's pathology-adjacent AI device count. This is proven technology; the question for a department is procurement, not validation of principle.

Bone marrow aspirate and trephine assessment. Considerably harder. Aspirate smears require the same multi-focal-plane handling as cytology; trephines require architectural assessment, cellularity estimation, and megakaryocyte morphology evaluation across a decalcified, often distorted specimen. As Courville noted in a 2026 *The Hematologist* technology review, most hematopathologists will not build these models — but they will be responsible for validating and monitoring them locally, which is a distinct and underappreciated skill set.

Flow cytometry analysis. Automated gating and unsupervised clustering of high-dimensional flow data is arguably the most under-exploited opportunity in the field. Manual gating is subjective, operator-dependent, and does not scale to the 20-plus-parameter panels now in routine use. Dimensionality-reduction and clustering methods handle this natively.

Measurable residual disease. MRD assessment by flow or by sequencing sits at the boundary of human perception. This is a quantitative, reproducible, high-stakes task — exactly the profile where algorithmic consistency has most to offer.

Lymphoma subclassification from histology. Weakly supervised models applied to lymph node whole-slide images can distinguish major categories, but the granularity that matters clinically — and that the WHO classification demands — remains dependent on immunophenotype and genetics.

Genomic and prognostic modelling. Machine learning applied to mutational profiles for risk stratification in myelodysplastic syndromes and AML has produced models that outperform established scoring systems in published cohorts. This is where haematology AI is most clinically advanced and least visually obvious.

3. Where to begin

If your laboratory has not adopted digital morphology, that is step one. Automated differential platforms are established technology with a regulatory track record and a workforce case that writes itself. This is the lowest-risk entry into machine-assisted haematology and it builds the digital habits everything else depends on.

If you already have digital morphology, look at flow. Automated gating pilots require no new instrument, no scanner, and no slide handling change. You can run unsupervised clustering against archived FCS files and compare to your manual gates retrospectively, at essentially zero clinical risk.

Do not begin with bone marrow trephine AI. It is the hardest imaging problem in the subspecialty, decalcification introduces enormous variability, and the evidence base is thin.

4. Your first 90 days

Weeks 1–4: inventory where your diagnostic data actually lives. In most departments, morphology sits in the LIS, flow in a separate analysis package, cytogenetics in a third system, and NGS output in a fourth. Any integrative AI ambition is blocked until these are queryable together. Mapping this is unglamorous and it is the real bottleneck.

Weeks 5–8: choose a retrospective, no-risk validation target. Archived flow files or archived digital differentials let you evaluate an algorithm against ground truth you already have, without touching a live case.

Weeks 9–12: define your ground truth carefully. In haematology, “correct” is frequently a consensus rather than a fact. If you are validating a blast-percentage algorithm, you need to know your own inter-observer variance on blast counting first — it is typically wider than people expect, and it sets the ceiling on any agreement statistic you will report.

5. Validation specifics

Stain and preparation variability is the dominant technical confounder. Romanowsky staining differs measurably between laboratories and between lots; models trained elsewhere frequently degrade on your smears. Validate across at least several stain batches.

Class imbalance is severe and clinically dangerous. Blasts, promyelocytes and abnormal lymphoid cells are rare relative to neutrophils. A model with excellent overall accuracy can be poor on precisely the cells that determine the diagnosis. Insist on per-class sensitivity, not aggregate accuracy.

Paediatric and adult marrows are not interchangeable. Normal paediatric marrow cellularity, haematogone populations and lineage proportions differ substantially. An algorithm validated on adults will over-call in children.

6. Failure modes to anticipate

Haematogones misread as blasts. The classic paediatric trap, and a well-documented weakness of both flow gating algorithms and morphology classifiers.

Post-therapy marrows. Regenerating marrow after induction chemotherapy looks abnormal to any model trained predominantly on diagnostic specimens. Dysplastic change, left shift and increased immaturity are expected findings, not disease.

Classification drift. WHO and ICC classification systems are being revised on a timescale faster than models are retrained. A model trained to a superseded classification will produce entities that no longer exist.

Over-reliance on integrative scores. Prognostic models built on mutational data reflect the treatment era of their training cohort. As therapy changes, the prognostic weights become stale even though the model does not report any uncertainty about this.

7. What to watch next

The integrative diagnosis is the prize. Models that take morphology, immunophenotype and genotype jointly and output a WHO-aligned classification with calibrated uncertainty would change hematopathology practice more than any single-slide algorithm. Several academic groups are building toward this, and the ASH AI subcommittee’s activity signals that professional-body guidance is coming. Watch also for automated MRD assessment, where regulatory and clinical trial pressure is strongest and where reproducibility gains translate directly into treatment decisions.

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

What is the per-class sensitivity for blasts, promyelocytes and abnormal lymphoid cells — not aggregate accuracy?

Was this validated on paediatric as well as adult marrows?

How does it perform on post-therapy and regenerating marrows?

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

Nazha et al., Artificial intelligence in hematology, *Blood*, 2025;146(19):2283–2292

ASH AI Subcommittee position

2

Courville, Digital microscopy and AI in bone marrow evaluation, *The Hematologist*, 2026;23(2)

Practical local validation framing

3

AI in cytopathological applications for cancer, *Eur J Med Res*, 2024

Includes marrow and peripheral blood WSI

4

FDA AI-Enabled Medical Device List, March 2026

Hematology device category

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.