Getting Started with AI in Clinical Microbiology

Path-iQ Getting Started Series · No. 18 of 20 · A practical primer for the practising pathologist

July 2026

1. The laboratory discipline with the clearest automation case

Clinical microbiology has a structural problem that AI addresses directly: it is the last major laboratory discipline still dependent on skilled human visual interpretation at high volume, and it faces severe workforce shortages. Chemistry autoverifies nearly everything. Haematology has automated differentials. Microbiology still has technologists reading plates and Gram stains by eye.

Total laboratory automation systems have brought specimen processing, plate streaking, incubation and digital imaging into a single workflow. Once plates are imaged rather than examined directly, the interpretation step becomes a computer vision problem — and that is where the field now sits.

2. Where the evidence is strongest

Culture plate screening and triage. Automated image analysis to screen urine cultures for growth and determine which require further workup has been evaluated in clinical laboratories and is among the most practically validated applications. The logic is compelling: the majority of urine cultures are negative or represent contamination, and technologist time spent confirming negatives is time unavailable for complex work.

Gram stain interpretation. The most active research area. A convolutional neural network developed at Harvard Medical School for blood culture Gram stain smears achieved 94.9% accuracy in distinguishing Gram-positive from Gram-negative bacteria. A transformer-based framework published in the *Journal of Clinical Microbiology* extended this to whole-slide Gram stain classification without requiring manual annotation — a significant practical advance, since annotation is the main cost in building such systems. Work published in late 2025 addressed classification into the six commonest morphological categories using 1,077 images drawn from multiple laboratories, microscopes and time periods, deliberately preserving the variability that defeats models trained on clean data.

A notable caveat from the literature: scanning microbiology slides is harder than scanning tissue, with uneven focal planes and background staining producing blurred regions on whole-slide images.

Virtual staining. Researchers at UCLA reported in *Science Advances* in early 2025 a method converting microscopic images of unstained bacteria into their Gram-stained equivalents using AI, bypassing chemical staining entirely. If this matures, it changes the pre-analytical workflow rather than the interpretation step.

Antimicrobial susceptibility. Rapid phenotypic systems have reduced identification and susceptibility turnaround from the conventional 48–72 hours to under 7 hours, with performance broadly equivalent or slightly superior to established phenotypic methods. Machine learning applied to susceptibility prediction from genomic data is a parallel and rapidly developing line.

Colony recognition. Work classifying fungal colonies — one study assembled 8,995 colony images across seven *Aspergillus* species — illustrates the approach for organisms where morphological identification remains standard.

3. Your first 90 days

Weeks 1–3: establish whether you have digital imaging. As with histology and scanning, the algorithm question is downstream of the imaging question. If your laboratory does not image plates or slides digitally, that is the project.

Weeks 4–8: start with urine culture screening. It is the highest-volume, most repetitive, most negative-dominated workflow in the laboratory, and the clinical risk of a well-validated screening step is manageable because positives are escalated rather than discarded.

Weeks 9–12: baseline your Gram stain concordance. Have multiple technologists read the same set of blood culture Gram stains and measure agreement. This is rarely done and typically reveals more variability than expected — particularly for Gram-variable organisms and mixed morphologies.

4. Validation specifics

Staining and microscope variability is the dominant technical confounder, and unusually the microbiology literature has confronted it directly. The most useful published work deliberately uses images from multiple laboratories, microscopes and periods, with variation in quality, scale and colour balance. When evaluating a tool, ask whether it was validated on multi-site data or on a single clean dataset; the difference predicts real-world performance.

Specimen type governs interpretation. A Gram stain from a sterile site — blood, cerebrospinal fluid, synovial or pleural fluid — carries different significance from a respiratory or wound specimen where colonisation is expected. An algorithm reporting organisms without specimen context is providing incomplete information.

Media and incubation conditions determine colony appearance. Models trained on one agar formulation may not transfer.

Local epidemiology and resistance patterns differ by institution and region. Susceptibility prediction models in particular are sensitive to the resistance mechanisms prevalent in the training population.

5. Failure modes to anticipate

Mixed cultures and polymicrobial specimens. Models trained on pure isolates handle mixtures poorly, and mixtures are common in real specimens.

Gram-variable organisms. *Clostridium*, *Bacillus* and older cultures stain unpredictably. These are precisely the organisms where morphological interpretation is hardest.

Contamination versus infection. The most consequential judgement in clinical microbiology, and one that depends on specimen quality, collection method and clinical context rather than on morphology alone.

Coagulase-negative staphylococci in a single blood culture bottle is the classic case.

Yeast and fungal elements. Under-represented in bacterial-focused training data and easily missed.

Autoverification of negatives. Automated release of negative results is the highest-efficiency and highest-risk application. A false-negative screen in a sterile-site specimen is a serious patient safety event, and the threshold must be set with that asymmetry explicitly in mind.

6. What to watch next

The realistic near-term picture is the fully automated negative pathway: specimens processed by total laboratory automation, plates imaged and read algorithmically around the clock, negatives autoverified into the LIS and discarded without technologist handling, and human expertise reserved for positives and complex work. Several components of this exist; the integration does not yet. Watch also for machine learning applied to susceptibility prediction from sequence data, which would compress the diagnostic timeline far more than faster phenotypic methods can.

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 multi-site data, or on a single laboratory’s clean dataset?

How does it perform on mixed cultures, Gram-variable organisms and fungal elements?

What is the false-negative rate for sterile-site specimens specifically?

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.

7. Key references

#	Source	Note
1	Automated characterisation of Gram-stained blood culture slides using a large-scale vision transformer, *J Clin Microbiol*, 2025;63:e01514-24	Annotation-free WSI classification
2	Ding et al., *iLABMED*, 2025	Harvard CNN, 94.9% Gram-positive vs negative accuracy; Accelerate Pheno <7h
3	Development of an AI Algorithm for Automatic Classification of Gram Stain Images, bioRxiv, 2025	1,077 multi-source images, six morphological categories
4	Faron et al., *J Clin Microbiol*, 2020;58:e01683-19	Automated image analysis for urine culture screening
5	Virtual Gram staining of label-free bacteria, *Science Advances*, 2025	AI-based stain-free imaging
6	Singhal et al., *Current Research in Biotechnology*, 2025;10:100351	Review of automation in clinical microbiology

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.