Getting Started with AI in Clinical Chemistry and Laboratory Medicine

Path-iQ Getting Started Series · No. 19 of 20 · A practical primer for the practising pathologist

July 2026

1. The discipline where AI has the most data and the least visibility

Clinical chemistry generates more discrete results than every other laboratory discipline combined. A single hospital laboratory produces millions of numeric results annually, each timestamped, each linked to a patient record, most with longitudinal history. This is the richest structured dataset in medicine, and it requires no scanning, no annotation and no image acquisition to use.

It is also the discipline where AI is least visible, because the applications are infrastructural rather than diagnostic. Nobody publishes a headline about an improved autoverification rule. But in terms of results affected per day, clinical chemistry is where machine learning touches the most patients.

2. Autoverification: the flagship application

Autoverification — automated release of results meeting defined criteria without technologist review — has existed for decades using rule-based logic. Machine learning improves on rules by learning patterns rules cannot express.

The published evidence is now substantial. A recent development study built an AI verification system for a clinical biochemistry laboratory using gradient-boosted trees, reporting an AUC-ROC of 0.970 (95% CI 0.954–0.973). Validation involved a double-blind comparison across 100,048 reports evaluated independently by laboratory staff and by the system, with staff assessment as the gold standard, followed by a turnaround-time comparison across 27,473 reports verified simultaneously by both. Earlier work in *Clinical Chemistry and Laboratory Medicine* had established the approach.

Reported figures of this kind should be read with attention to one number in particular: the false-negative rate, meaning results the system passed that staff would have held. In autoverification, that is the only error that reaches the patient, and it is the metric that should govern threshold setting.

3. The task map

Autoverification and error detection. As above. Detection of pre-analytical errors — haemolysis, contamination from an infusion line, wrong-blood-in-tube, specimen mix-up — is a related and high-value application, since delta-check logic based on fixed rules is crude and machine learning can use much richer context.

Result interpretation and clinical decision support. Flagging patterns across multiple analytes that suggest a specific condition, and prompting reflex testing. A comprehensive review of machine learning in laboratory medicine categorised applications across twelve topics including autoverification, clinical decision support, error detection, disease screening, estimation and prediction, with input dataset sizes

ranging from a handful of samples to 25 million.

Reference interval derivation. Indirect methods deriving reference intervals from accumulated patient data rather than recruited healthy volunteers. Statistically demanding, and machine learning approaches handle the mixture-model problem well.

Quality control. Patient-based real-time quality control, using moving averages of patient results to detect analytical drift between conventional QC events. This is one of the most practically valuable and under-adopted applications in the discipline.

Demand management. Predicting which requested tests are unlikely to change management, and identifying duplicate or excessively frequent ordering.

Workload and inventory forecasting. Predicting specimen volumes, staffing needs and reagent consumption.

4. Your first 90 days

Weeks 1–3: measure your current autoverification rate and your manual review burden. In many laboratories a substantial proportion of manually reviewed results are released unchanged, which quantifies the opportunity directly.

Weeks 4–8: start with patient-based real-time quality control. It requires no new instrumentation, uses data you already generate, carries no direct patient risk during evaluation, and detects analytical problems that scheduled QC misses. It is the single best entry point in this discipline.

Weeks 9–12: run a retrospective autoverification evaluation. Take a large archive of results with their manual review outcomes and evaluate a model against them. Because ground truth already exists in your LIS, this can be done at essentially zero clinical risk and with sample sizes that make the statistics meaningful.

5. Validation specifics

Sample size expectations are different here. In imaging pathology a 200-case validation may be adequate. In autoverification, where you are trying to characterise an error rate that must be very low, you need tens of thousands of results to bound it usefully. Design accordingly.

Population and case mix govern everything. A model trained in a tertiary centre with high acuity will behave differently in primary care. Paediatric, neonatal, obstetric, oncology and dialysis populations all have distinct result distributions.

Analyser, method and reagent lot changes shift result distributions measurably, and any model must be monitored across these transitions. Method changes are the commonest cause of silent model degradation in chemistry.

Reference interval and unit consistency across the training and deployment environments must be confirmed explicitly — an avoidable but recurring source of error.

6. Failure modes to anticipate

Rare critical results. Extreme values are rare by definition and therefore under-represented in training. A model with excellent overall performance may be worst on precisely the results that require immediate action. Demand performance stratified by result magnitude.

Interference. Haemolysis, icterus, lipaemia, paraproteins, biotin and macro-analytes produce spurious results. A model that has not learned these will pass them.

Drift after method change. The most common real-world failure. Model performance must be re-verified after any assay, analyser or reagent change, and this must be written into the quality management system.

Automation of judgement. Deciding whether a result is plausible for a given patient sometimes requires clinical information the laboratory does not have. Keep escalation pathways open and staffed.

Over-restriction of testing. Demand management models that reduce test ordering can reduce appropriate testing along with inappropriate testing. Monitor clinical outcomes, not just volume reduction.

7. What to watch next

Two things. First, the extension of longitudinal modelling — using a patient's own result history rather than population reference intervals to determine whether a value is abnormal for them. This is a better statistical framework than fixed reference intervals and the data to support it already exists in every LIS. Second, integration of laboratory data into hospital-wide predictive models for deterioration, sepsis and acute kidney injury. Laboratory results are the highest-signal inputs to these models, which means laboratory professionals should be involved in their governance — and frequently are not.

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 false-negative rate — results passed that staff would have held?

Is performance stratified by result magnitude, including critical values?

What is the revalidation requirement after an assay, analyser or reagent lot 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

Zhu et al., *iLABMED*, 2025

XGBoost autoverification; AUC-ROC 0.970; 100,048 reports double-blind

2

Wang et al., *Clin Chem Lab Med*, 2020;59:883–891

Machine learning autoverification in clinical biochemistry

3

Advancing Laboratory Medicine Practice With Machine Learning, 2025

Review; twelve application categories

4

Patient-based real-time quality control literature

Moving average QC methodology

5

FDA AI-Enabled Medical Device List, March 2026

Clinical chemistry 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.