Getting Started with AI in Breast Pathology

Path-iQ Getting Started Series · No. 10 of 20 · A practical primer for the practising pathologist

July 2026

1. The most quantified subspecialty in pathology

Breast pathology has been converting morphology into numbers for longer than any other subspecialty. Nottingham grade decomposes into three scored components. ER, PR, HER2 and Ki-67 are all reported as percentages or scored categories. Tumour size, node status and margin distance are measurements. Multigene assays convert all of it into a recurrence score.

This makes breast pathology the natural home of quantitative image analysis, and it is where automated scoring first entered routine practice — well before “AI” was the framing.

It is also where the reproducibility problems are best documented, which is why the field moved. HER2-low classification is the current exemplar: with the arrival of HER2-directed antibody-drug conjugates effective in tumours expressing HER2 at levels just above background, the clinically consequential distinction became the one pathologists agree on least. Published work has shown low interobserver agreement even among subspecialised breast pathologists evaluating HER2-low. A therapeutic decision now rests on a call that experts make inconsistently — an unusually clean argument for algorithmic assistance.

2. The regulatory position

Breast pathology gained its own prognostic milestone in May 2026, when ArteraAI Breast received FDA clearance for use in early-stage, hormone-receptor-positive, HER2-negative invasive breast cancer — the first FDA-cleared digital-pathology-based risk stratification tool for breast cancer. This followed the same developer’s prostate authorisation and confirms the prognostic lane as a durable regulatory category rather than a one-off.

On the detection side, Paige Breast and Paige Breast Lymph Node address neoplasm identification and nodal metastasis detection, the latter reported at over 98% sensitivity. Commercial quantification modules for HER2, Ki-67 and ER/PR from several vendors are designed to run without per-laboratory manual tuning, which is the practical requirement for multi-site consistency.

3. The task map

Lymph node metastasis detection. The canonical computational pathology benchmark, established by the CAMELYON challenges. Micrometastasis and isolated tumour cell detection is exactly the needle-in-haystack task where sustained human attention fails and algorithmic screening excels.

Biomarker quantification. ER and PR percentage, Ki-67 index, HER2 scoring. Ki-67 in particular has such poor interlaboratory reproducibility that international working groups have repeatedly cautioned against threshold-based clinical use — an automated, calibrated measurement addresses the root cause.

HER2 and HER2-low. Beyond conventional scoring, models have been developed to infer HER2 expression from H&E alone, including identification of HER2-low cases. Treat these as triage tools; the therapeutic decision requires the assay.

Mitotic counting. One of the three Nottingham components, dependent on field selection and hotspot identification, and demonstrably variable. Automated mitotic figure detection with hotspot localisation and density calculation is among the most solid quantification applications available.

Tumour-infiltrating lymphocytes. Prognostic and predictive, particularly in triple-negative and HER2-positive disease, and scored by visual estimation with wide variability. TILs quantification is a well-defined computational task with an international scoring standard to calibrate against.

DCIS and margin assessment. Detection of ductal carcinoma in situ and automated margin distance measurement, both with direct surgical consequences.

4. Your first 90 days

Weeks 1–2: pick the biomarker with your worst reproducibility. In most departments this is Ki-67. Run a simple exercise: have three pathologists score the same 30 cases. Publish the spread internally. This single exercise generates more institutional support for quantitative AI than any vendor presentation.

Weeks 3–6: evaluate an immunohistochemistry quantification module. These are the most mature, lowest-risk tools available. The output is a number you can compare directly against your manual score, ground truth is definable, and there is no diagnostic risk in shadow mode.

Weeks 7–12: consider nodal screening. Metastasis detection on sentinel nodes is high-volume, has clear ground truth on levels and cytokeratin stains, and delivers a genuine safety benefit. Log every discordance and review with cytokeratin immunohistochemistry.

5. Validation specifics

Pre-analytical variables dominate breast biomarker performance. Cold ischaemia time, fixation duration and antigen retrieval protocol affect ER, PR, HER2 and Ki-67 staining intensity directly. An algorithm reading intensity will faithfully report your pre-analytical problems as biology. Confirm your fixation compliance before blaming an algorithm.

Assay platform matters. HER2 immunohistochemistry differs between manufacturers’ antibodies and detection systems. A model calibrated on one platform will not transfer without revalidation.

Biopsy versus excision. Core biopsies are the material on which treatment decisions are made and are the appropriate validation substrate, but many models are trained on resections.

Neoadjuvant-treated specimens are a distinct problem. Residual cancer burden assessment after neoadjuvant chemotherapy involves scattered residual tumour cells in fibrotic beds — morphologically unlike treatment-naive tumour and under-represented in training data.

6. Failure modes to anticipate

Lobular carcinoma. Single-file infiltration with minimal desmoplasia is systematically harder for detection algorithms than ductal carcinoma, and it is also where human detection of nodal metastases is weakest. The two failures compound.

Benign mimics. Sclerosing adenosis, radial scars and microglandular adenosis generate false positives.

Heterogeneous HER2 expression. Intratumoural heterogeneity means whole-slide averaging can misrepresent a focally amplified tumour.

Isolated tumour cells. Detection is achievable; the harder question is whether flagging every isolated tumour cell cluster changes management, or simply generates work. Decide your reporting policy before deploying.

7. What to watch next

The prognostic category is expanding fastest. Tools that predict treatment benefit — chemotherapy benefit, endocrine therapy benefit, response to antibody-drug conjugates — from the slide plus clinical data are the direction of travel, and breast is the tumour type with the richest trial data to train on. Watch also for H&E-based biomarker inference maturing to the point of formal regulatory submission; if a model can reliably identify HER2-low from H&E, the reflex testing algorithm in every breast unit changes.

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

Which immunohistochemistry clone and detection platform was this calibrated on?

Was it validated on core biopsies, which is where treatment decisions are made?

How does it perform on lobular carcinoma and on post-neoadjuvant specimens?

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

ArteraAI Breast FDA clearance, May 2026

First FDA-cleared digital pathology risk stratification tool in breast cancer

2

Turashvili et al., *J Clin Pathol*, 2024;77:815–821

Low interobserver agreement in HER2-low

3

AI and digital tools in cancer pathology, *Lancet Digital Health*, 2025

HER2, Ki-67, PD-L1 image analysis review

4

AI’s impact on breast cancer pathology: a literature review

Paige Breast Lymph Node >98% sensitivity; commercial quantification modules

5

CAMELYON16/17 challenge publications

Benchmark for nodal metastasis detection

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.