Getting Started with AI in Head and Neck Pathology

Path-iQ Getting Started Series · No. 11 of 20 · A practical primer for the practising pathologist

July 2026

1. A subspecialty defined by anatomical complexity and diagnostic rarity

Head and neck pathology is unusually difficult to automate for a structural reason: it encompasses more distinct anatomical sites, tissue types and tumour entities than almost any other subspecialty, while individual entity volumes remain low. Salivary gland tumours alone comprise more than twenty recognised types, many morphologically overlapping and several defined by specific gene fusions.

This inverts the usual AI calculus. Where breast and prostate offer high volume and narrow differentials — ideal training conditions — head and neck offers the opposite. The realistic near-term role for AI here is therefore not autonomous classification but decision support, quantification and molecular triage.

2. Where the strongest evidence sits: molecular classification

The most convincing head and neck AI result is not an image model. Jurmeister and colleagues published DNA methylation-based classification of sinonasal tumours in *Nature Communications* in 2022, applying the same machine learning approach that transformed neuropathology to a site notorious for diagnostic difficulty. Sinonasal tumours are rare, morphologically ambiguous, and frequently misclassified; a methylation classifier provides an objective anchor where morphology and immunohistochemistry frequently do not resolve the question.

This is the pattern worth internalising for head and neck: where morphology is genuinely ambiguous and entity definitions are molecular, methylation-based and sequencing-based classifiers deliver more than image models do.

3. The task map

HPV status in oropharyngeal carcinoma. The single most consequential distinction in the subspecialty, driving staging and treatment de-escalation. Currently determined by p16 immunohistochemistry with HPV-specific testing in defined circumstances. Deep learning models predicting HPV status directly from H&E have been reported with reasonable performance and are an appealing triage application — though p16 immunohistochemistry is cheap and available, which limits the practical gain.

Oral epithelial dysplasia grading. Among the least reproducible grading systems in pathology, with direct implications for excision versus surveillance. Binary and three-tier systems both show poor agreement. A quantitative, reproducible measure would be genuinely valuable.

Oral squamous cell carcinoma: depth of invasion and pattern of invasion. Depth of invasion entered T staging in the eighth edition of AJCC and is a measurement — precisely the kind of task where automated measurement removes variability. Worst pattern of invasion is prognostic and subjectively assessed.

Perineural and lymphovascular invasion. Focal, easily missed, prognostically important, and a natural screening application.

Margin assessment. Head and neck resections have complex three-dimensional margins with significant implications for adjuvant therapy. Automated margin distance measurement on scanned sections is practical and useful.

Salivary gland tumour classification. Realistically a decision-support and differential-generation problem rather than an automation target, given entity rarity.

Thyroid. Fine-needle aspiration is discussed in the cytopathology primer; on the histology side, the diagnosis of encapsulated follicular-patterned lesions — follicular adenoma, NIFTP, follicular variant papillary carcinoma, follicular carcinoma — is the classic reproducibility problem, hinging on capsular and vascular invasion assessment that is both tedious and inconsistent.

4. Your first 90 days

Weeks 1–3: identify your genuine reproducibility pain point. For most departments this is oral epithelial dysplasia grading or thyroid follicular lesion assessment. Measure it. Circulate 30 cases among your reporters and quantify agreement.

Weeks 4–8: start with a measurement task, not a classification task. Depth of invasion, margin distance and capsular invasion assessment are measurable, verifiable and low-risk. Classification of rare entities is not a sensible first project in this subspecialty.

Weeks 9–12: audit your molecular pathway. Determine what proportion of your ambiguous sinonasal and salivary cases currently receive molecular characterisation, and what the referral route is. In head and neck, expanding molecular access will usually deliver more diagnostic improvement than any imaging algorithm you could deploy.

5. Validation specifics

Site heterogeneity is the defining validation challenge. A model validated on oral cavity squamous carcinoma has not been validated for larynx, sinonasal tract or oropharynx, even though the tumour type is nominally the same. Insist on site-specific performance data.

Decalcification affects any specimen involving bone — mandibular resections, temporal bone — and degrades both morphology and immunohistochemistry. Models rarely account for this.

Prior radiotherapy is common in this population and produces atypia that mimics recurrence. Post-treatment neck dissections are a distinct and under-represented specimen category.

Small biopsies from mucosal sites are frequently crushed, tangentially oriented and cauterised. This is normal material in head and neck practice and unusual in most training sets.

6. Failure modes to anticipate

Reactive and reparative atypia. Ulcerated mucosa, granulation tissue and post-biopsy change generate false positives for dysplasia and carcinoma.

Lymphoid tissue. Waldeyer's ring, cervical nodes and tonsillar tissue are lymphoid-rich, and models trained on epithelial malignancy handle dense lymphoid infiltrates poorly. Basaloid and non-keratinising carcinoma within tonsillar crypts is a specific trap.

Rare entity default. When shown an entity outside its training set, a classifier assigns the nearest class rather than declaring uncertainty. In a subspecialty defined by rare entities, this is the dominant risk. Insist on models that abstain.

Fusion-defined tumours. Secretory carcinoma, mucoepidermoid carcinoma and adenoid cystic carcinoma are defined by molecular events with variable morphology. Image models will not reliably capture this.

7. What to watch next

Two developments matter. First, extension of methylation-based classification beyond sinonasal to salivary and odontogenic tumours — the method has proven generalisable and the diagnostic need is acute. Second, pan-tissue foundation models with genuine abstention behaviour, which would allow a rare-entity subspecialty to benefit from AI screening without the classifier confidently mislabelling the unusual case. Until abstention is standard, head and neck should treat classification outputs with more scepticism than higher-volume subspecialties need to.

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

Is performance reported by anatomical subsite, or pooled across head and neck?

How does the model behave on decalcified and post-radiotherapy specimens?

Does the classifier abstain on entities outside its training set, or assign the nearest class?

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

Jurmeister et al., *Nature Communications*, 2022;13:7148

DNA methylation-based classification of sinonasal tumours

2

WHO Classification of Head and Neck Tumours, 5th edition

Entity definitions and molecular criteria

3

AJCC 8th edition staging

Depth of invasion as a staging parameter

4

Benchmarking foundation models, *Nature Biomedical Engineering*, 2025

Generalisability across tissue types

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.