Getting Started with AI in Bone and Soft Tissue Pathology

Path-iQ Getting Started Series · No. 14 of 20 · A practical primer for the practising pathologist

July 2026

1. The rarity problem, stated plainly

Soft tissue and bone tumours comprise more than a hundred recognised entities, the great majority of them rare, many of them defined by specific genetic events, and collectively accounting for around 1% of adult malignancies. This is the hardest possible environment for supervised machine learning, which depends on seeing many examples of each class.

It is also the subspecialty where diagnostic error carries some of the highest costs, since treatment ranges from observation to limb-sparing surgery to amputation, and second-opinion discordance rates at referral centres are substantial.

The resolution of this tension is the defining feature of AI in sarcoma pathology: the useful applications are prognostic and molecular, not diagnostic classification.

2. Where the evidence is strongest

Prognostication. The FNCLCC grade is the primary prognostic factor in soft tissue sarcoma, and it is well recognised that it does not fully capture high-risk patients. Deep learning has been applied directly to this gap. Foersch and colleagues reported deep learning for diagnosis and survival prediction in soft tissue sarcoma in *Annals of Oncology* in 2021. More recently, Michot and colleagues published a model in *Scientific Reports* in 2025 predicting metastatic relapse-free survival from digital pathology of both tumour and margin areas — an interesting design choice, since the margin is tissue pathologists examine for adequacy rather than for prognostic information.

The pattern is consistent with the wider literature: image-based prognostication in sarcoma adds information beyond grade, which is a genuinely useful contribution and does not require the model to name the entity.

Molecular classification. Sarcoma classification by DNA methylation profiling has been established as a reliable approach, following the same methodology that reorganised neuropathology. For a subspecialty defined by ambiguous morphology and specific genetic events, this is arguably more valuable than any image model. Where morphology cannot separate entities, methylation frequently can.

3. The task map

Grading components. FNCLCC grade combines differentiation, mitotic count and necrosis percentage. Two of these three are quantitative measurements performed by estimation. Automated mitotic counting and necrosis percentage quantification are directly applicable and address known variability.

Treatment response assessment. Percentage necrosis after neoadjuvant chemotherapy is the key prognostic determinant in osteosarcoma and Ewing sarcoma, conventionally assessed by mapping resection specimens and estimating necrosis — an enormously tedious, poorly reproducible task, and one of the strongest quantification targets anywhere in pathology.

Margin assessment. Distance to margin determines adjuvant radiotherapy decisions.

Bone tumour matrix characterisation. Distinguishing osteoid, chondroid and fibrous matrix, and assessing permeative growth.

Triage and screening. Flagging cases that warrant subspecialist referral. Given that most sarcomas are first seen by a generalist, a tool that reliably says “this is not routine” has real value even without naming the entity.

4. Your first 90 days

Weeks 1–3: choose necrosis quantification. For any centre treating bone sarcomas, this is the clearest target: it is a percentage, it is currently estimated across many slides, it determines prognosis and adjuvant decisions, and ground truth is definable by careful manual annotation.

Weeks 4–8: baseline your mitotic counting. Have several pathologists count mitoses in defined fields on the same cases. The spread will justify the project.

Weeks 9–12: audit your molecular access. Determine how many of your ambiguous spindle cell and round cell lesions receive fusion testing or methylation profiling, and what the turnaround is. In this subspecialty, expanding molecular characterisation will improve diagnostic accuracy more than any imaging tool currently available.

5. Validation specifics

Decalcification is the dominant technical variable in bone pathology. It alters morphology, degrades nucleic acids, and severely affects immunohistochemistry. Decalcification protocols differ between laboratories, and models trained on one will not transfer reliably. Always ask what protocol the training material used.

Entity representation must be interrogated case by case. A model reporting excellent performance across “sarcoma” may have been trained overwhelmingly on liposarcoma, leiomyosarcoma and undifferentiated pleomorphic sarcoma, which together dominate incidence. Its performance on synovial sarcoma or epithelioid sarcoma may be untested.

Core biopsy versus resection is critical. Sarcoma treatment planning rests on core biopsy, which shows limited tissue with sampling bias toward viable, accessible areas. Most training data is resection-derived.

6. Failure modes to anticipate

Benign mimics. Nodular fasciitis, myositis ossificans, proliferative fasciitis and desmoid fibromatosis are benign lesions that mimic sarcoma histologically and are classic causes of over-diagnosis. They are exactly the lesions where an over-confident model is dangerous.

Post-treatment change. Radiotherapy-induced atypia in mesenchymal tissue mimics sarcoma. Post-radiation sarcoma versus radiation atypia is a genuinely difficult call.

Dedifferentiation. Dedifferentiated liposarcoma and dedifferentiated chondrosarcoma show abrupt transitions that whole-slide averaging misrepresents.

The confident wrong answer on a rare entity. In a subspecialty where most entities are rare, a classifier without abstention behaviour will produce a confident, plausible and incorrect label. This is the central risk and it should govern procurement: prefer tools that decline to answer.

7. What to watch next

Methylation-based sarcoma classification becoming routinely accessible would do more for diagnostic accuracy in this subspecialty than any image model, and its adoption curve is the thing to follow. On the imaging side, watch for prognostic models being validated prospectively within sarcoma clinical trials — the Michot approach of extracting prognostic signal from tissue we already examine but do not currently interrogate is a genuinely novel contribution, and prospective validation is the next step it requires.

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 decalcification protocol was the training material processed with?

Which specific entities are represented in training, and at what case numbers?

Does the model abstain on entities it has not seen, and how is that surfaced?

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

Michot et al., *Scientific Reports*, 2025;15:38534

DL prognostic prediction from tumour and margin areas in STS

2

Foersch et al., *Annals of Oncology*, 2021;32:1178–1187

DL for diagnosis and survival prediction in soft tissue sarcoma

3

Koelsche et al.

Sarcoma classification by DNA methylation profiling

4

WHO Classification of Soft Tissue and Bone Tumours, 5th edition

Entity definitions and molecular criteria

5

FNCLCC grading system

Prognostic framework and its known limitations

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.