MedCrave
Step into the World of Research

Journal of Stem Cell Research & Therapeutics

Narrative Review

Open Access

CrossMark

# Digital pathology and artificial intelligence in cancer stem cell research: from quantification to spatial mapping

## Abstract

Digital pathology and artificial intelligence (AI) can quantify cancer stem cell (CSC)-associated tissue phenotypes and examine their spatial organization, but the resulting measurements support different biological claims. This narrative review examines five purposively selected application studies, together with enabling technologies, functional experiments and validation guidance. Targeted source searches documented in the original manuscript on 26 October 2026 were supplemented by source verification on 6 November 2026. The synthesis compares cohort composition, reference labels, evaluation units and independence of testing; it does not estimate pooled accuracy or claim exhaustive coverage. The applications include exploratory liver image classification, nuclear morphometry associated with transcriptomic stemness, CD133-containing multiplex phenotyping, a uveal melanoma prognostic signature and virtual CD44 staining. Here, spatial mapping means localization and quantitative analysis of measured or inferred phenotypes at a stated cellular, neighborhood or regional scale, rather than direct identification of a functional CSC niche. Patient-separated internal testing, external cohort evaluation and functional corroboration answer distinct questions. The proposed framework links each claim to its output, reference standard, evaluation population and permitted interpretation, complementing existing assay-validation and reporting guidance. It is an unvalidated planning synthesis, not a diagnostic algorithm. Current evidence supports defined research measurements and hypothesis generation; claims about self-renewal, tumor initiation or treatment benefit require additional functional or clinical evaluation.

**Keywords:** cancer stem cells, digital pathology, artificial intelligence, immunohistochemistry, CD44, CD133, spatial analysis, validation

Volume 11 Issue 1 - 2026

Gamze Mercan
Hacettepe University, Ankara, Turkey

**Correspondence:** Assoc. Prof. Dr. Gamze Mercan, Hacettepe University, Ankara, Turkey

**Received:** October 27, 2026 | **Published:** November 23, 2026

## Introduction

Cancer stem cell (CSC) research asks which tumor cells can sustain growth, generate heterogeneous progeny or contribute to disease persistence. Histopathology asks where those cells and their associated phenotypes occur within tissue. Bringing these questions together requires more than assigning a stem cell label to a brown or fluorescent signal. The tumorigenic breast cancer population described by Al-Hajj and colleagues was supported by transplantation experiments as well as a surface phenotype.[1] Likewise, the mammary stem and progenitor evidence reported by Ginestier and colleagues combined aldehyde dehydrogenase activity with biological assays.[2] These studies illustrate the evidential distance between measuring a marker and demonstrating a cellular function.

Digital pathology can help bridge part of that distance by retaining tissue architecture while making observations countable. Artificial intelligence (AI) can segment nuclei, classify image regions or predict a molecular label. Each task creates a different output and needs a different reference standard. This review examines that distinction through concrete studies rather than treating all applications as automated cancer stem cell detection. The emphasis is on research questions that a pathology laboratory can define, measure and independently test.

## Scope and literature approach

This is a selective narrative review organized around the question of what tissue-image measurements can establish about CSC-associated phenotypes. The original manuscript documents targeted web searches and source checks completed on 26 October 2026, using combinations of cancer stem cells, stemness, digital pathology, histopathology, deep learning, CD44, CD133, multiplex immunofluorescence and virtual staining. PubMed and PubMed Central records, original publisher pages and institutional repositories were used to check primary findings and bibliographic details. These sources served as discovery or verification resources; the work was not a reproducible, exhaustive search of each database. An auditable count of records retrieved, screened and excluded is unavailable, and no retrospective screening totals are supplied.

During revision on 6 November 2026, targeted web searches rechecked the five application reports and the guidance used to position the framework. Examples of searches actually used were: “Automated cancer stem cell recognition” “0.936”; “Automated Nuclear Segmentation” Blocker 49 stemness; “Multiplex immunofluorescence” “553” Peng CD133; “Multi-omics integrative analysis” “stemness-associated” uveal; and “Accelerating histopathology workflows” “VirtualMultiplexer” cohort. Exact article titles or DOIs were then used to resolve source records. Search results were checked against primary publications or author-institution repositories where accessible. The revision checks focused on the interpretation and reporting of the existing evidence set and do not establish comprehensive coverage through the revision date.

Selection for the application comparison required a primary report connecting histological images with a CSC-associated marker, an explicit molecular stemness construct, or virtual generation of a relevant marker. Five reports were retained as illustrative applications because they represent different outputs: morphological



©2026 Mercan. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

classification,[24] morphometric association,[25] multiplex spatial phenotyping,[26] pathomic prognosis[27] and virtual staining.[32] The conference report[24] was retained as early feasibility evidence and is identified as such. General image-analysis methods without a CSC-associated endpoint were discussed separately as enabling technology. Foundational functional experiments and original validation or reporting guidance were included to evaluate the interpretation of these outputs. Studies without a tissue-image component, such as stand-alone compound-screening AI, were outside the application comparison. This purposive selection supports methodological comparison, not a complete inventory of the field.

For the application reports, the comparison records tumor or tissue setting, patients separately from image units where verifiable, imaging modality, computational output, reference measurement or label, evaluation design, and the boundary of the supported claim. An internal held-out subset and patient-level cross-validation are distinguished from an external cohort. External molecular corroboration is distinguished from external validation of an image-based prognostic model. Unverified counts or procedures are identified as unavailable rather than inferred from the nominal size of a public dataset. Where only an abstract or source record could be verified, no additional full-text-only methodological detail was imputed. No duplicate screening, formal risk-of-bias scoring or meta-analysis was performed. The comparison is a qualitative methodological appraisal, and the conceptual figures and framework contain no newly generated biological observations.

Throughout the synthesis, direct application evidence is separated from enabling technology. A general segmentation algorithm can support a future marker study without validating that marker, and a bulk molecular label does not provide a cell-level reference. The same distinction governs terminology: marker-positive refers to a specified assay, inferred stemness refers to a computational construct, and functional CSC evidence refers to an appropriate biological experiment. These terms define the scope of the review rather than implying equivalence among the selected studies.

## Define the biological endpoint before training a model

The first endpoint is immunophenotype: a cell or tissue region meets a specified staining criterion. Even that definition needs molecular precision. Kemper and colleagues showed that AC133 epitope recognition could decline during differentiation without corresponding loss of CD133 protein.[3] A model trained against AC133 annotations therefore learns an assay-dependent label, not a universal CD133-negative state. Marcato and colleagues further demonstrated an important contribution of ALDH1A3 to ALDH activity in the breast cancer systems they examined.[4] ALDH1A1 staining, ALDH1A3 staining and viable-cell ALDH activity should consequently remain separate labels.

The second endpoint is cellular state or molecular association. Liu and colleagues described epithelial-like and mesenchymal-like breast cancer stem cell states with different marker patterns and tissue distributions.[5] This provides a reason to examine location as well as abundance, but it does not supply a fixed morphological definition for every tumor. The third endpoint is function, established through experiments appropriate to the claim. The fourth is clinical use, such as improving a treatment decision. In this review, these endpoints are treated as distinct levels of evidence rather than interchangeable descriptions of model performance.

## Functional stemness is a dynamic experimental question

Lineage tracing illustrates why functional evidence cannot be read directly from a static image. Schepers and colleagues followed Lgr5-positive cells within mouse intestinal adenomas and showed their contribution to continuing tumor growth and the generation of other adenoma cell types.[6] The relevant observation was persistence and production of descendants over time. A single tissue section, even when segmented perfectly, contains neither the temporal sequence nor the experimental labeling that made that conclusion possible. Image analysis can locate a candidate phenotype, while lineage tracing tests what that population actually does in the chosen model.

Human colorectal organoid work adds a further complication. Shimokawa and colleagues demonstrated self-renewal and differentiation of LGR5-positive tumor cells and described re-emergence of the LGR5-positive population after selective ablation.[7] In a distinct mouse colorectal cancer setting, de Sousa e Melo and colleagues found different roles for Lgr5-positive cells in primary growth and metastatic disease.[8] These experiments should not be flattened into a universal statement about all cancers. They establish that cellular state, experimental context and the endpoint being tested can change the meaning of a marker. For digital pathology, a label derived at diagnosis may consequently answer a different question from one derived after treatment.

The resulting research opportunity is longitudinal pairing. An image-derived phenotype can be measured before and after a defined perturbation, with functional readouts collected independently. The image is then one component of an experimental argument rather than the sole reference for it. A change in predicted positivity should be checked against actual staining and specimen quality before being attributed to cellular conversion. When successive biopsies come from different locations, regional sampling must also be distinguished from temporal change. These are proposed design principles; the present review does not report a longitudinal cohort.

## Antibody choice can redefine the computational target

The comparative glioblastoma study by Hermansen and colleagues is a useful warning for annotation design. Four CD133 antibody clones produced discordant tissue staining patterns.[9] A computational model trained on one of those assays can be accurate relative to its own labels while remaining non-equivalent to a model trained on another assay. This is not solved by increasing the number of annotations alone. The target definition includes the clone, preparation method, expected cellular localization and the criteria by which a pathologist accepts the signal. Training labels should retain those details.

CD44 raises a related question about isoforms. Boman and colleagues used an epitope-directed antibody panel to examine CD44 variants in normal colonic mucosa and carcinoma, reporting CD44v8–10 expression in the stem cell niche and malignant tissue.[10] A pan-CD44 image label does not contain the same molecular information as a variant-directed label. Similarly, an algorithm that separates positive from negative tissue cannot recover an isoform distinction that the stain itself does not make. Digital analysis is therefore best specified as quantification of a defined assay rather than quantification of an unrestricted biological category.

Recent experimental results also separate a marker from a causal regulator. In the Caco2 model examined by Kholodenko and

Copyright: ©2026 Mercan.


colleagues, CD133 knockout did not reproduce the molecular and phenotypic effects associated with TRIM28 knockout, although TRIM28 loss reduced CD133 protein abundance.[11] The appropriate interpretation is model-specific. It supports asking whether a marker accompanies a state, contributes to that state, or is merely affected by another regulator. An image predictor of CD133 expression does not settle that causal question. This distinction is especially important when a computational association is used to motivate a therapeutic target.

Figure 1 separates measured phenotypes, inferred phenotypes and experimentally tested functions. The purpose of the distinction is practical: a study can be useful at any of these levels if its conclusion remains matched to its reference standard.

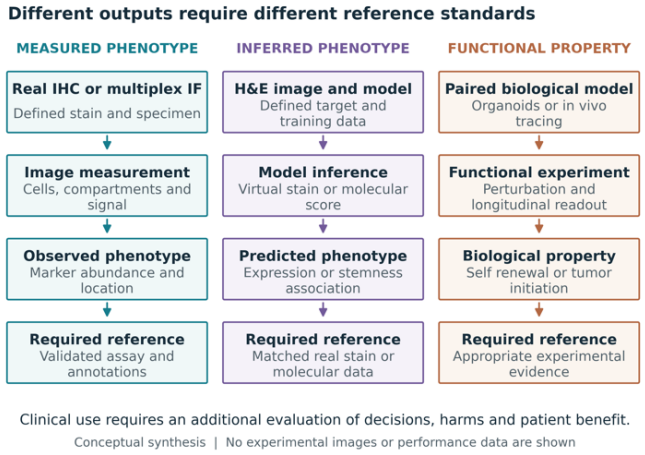


**Figure 1** Distinguishing measured phenotypes from model inferences and functional properties.

Original conceptual synthesis distinguishing observed staining, model inference and experimentally tested function.[1-12] Each track requires its own reference standard, and clinical utility requires separate evaluation. A measured phenotype denotes the specified assay, not universal CSC identity; a predicted phenotype remains an inference. IHC, immunohistochemistry; IF, immunofluorescence; H&E, hematoxylin and eosin. No experimental images, patient observations or performance data are shown.

## From a scanned slide to a defensible measurement

Specimen quality is part of the model specification

Specimen history should be recorded before the first image is selected for model development. BRISQ provides a framework for reporting collection, processing and storage conditions that can influence biospecimen research.[13] In a retrospective marker study, missing metadata is not simply an administrative inconvenience. Cases from one period or institution may differ in both clinical characteristics and tissue processing. If the model is trained against an outcome associated with those differences, an apparent biological signal may partly reflect how the specimens were handled. Recording unknown values allows that uncertainty to be assessed rather than silently removed.

Cold ischemic time has been experimentally investigated for established breast cancer biomarkers, including estrogen receptor, progesterone receptor and HER2.[14] Those results are adjacent evidence about preanalytical sensitivity; they do not supply validated acceptance limits for every CSC-associated antibody. A laboratory should resist importing a fixation rule from an unrelated analyte without verification. For image studies, the practical implications are to preserve specimen-type and processing metadata, inspect staining quality, and evaluate whether excluding poorly documented cases changes the conclusion. A computational pipeline should not convert unknown handling into an assumed standard condition.

Controls also determine the reliability of the labels used for training. Recommendations from the international expert panel on negative controls distinguish tissue controls from reagent controls and emphasize selection appropriate to the purpose of the assay.[15] An omitted-primary-antibody control can reveal some background processes but does not, by itself, establish the specificity of the primary antibody. Within a digital workflow, control review should precede annotation and scoring. A technically failed slide should not enter the training set as a confidently negative biological example. Recording the reason for exclusion helps readers distinguish quality selection from selective removal of difficult cases.

## Segmentation and tumor-cell identification

A typical workflow starts with tissue selection, image quality assessment and a defined region of interest. QuPath provides an open-source environment for whole-slide analysis, tumor identification and biomarker evaluation.[16] These functions make measurement accessible, but software availability does not establish assay validity. The laboratory still needs to specify the stain, compartment and cell population that constitute its endpoint.

Nuclear segmentation is an enabling step rather than a biological conclusion. HoVer-Net was developed to segment and classify nuclei in histology images, including the separation of clustered nuclei.[17] A nuclear boundary, however, is not a validated membrane or cytoplasmic boundary. For CD44 or CD133 research, signal assignment near adjacent cells requires particular attention. An apparently small segmentation error can change which cell is labeled positive and can also distort an inferred cell-to-cell distance. Inspection should therefore include crowded, weakly stained and morphologically ambiguous regions, not only visually clean fields.

Whole-cell and nuclear segmentation should be evaluated as separate choices. Cellpose 2.0 demonstrates a practical human-in-the-loop approach for adapting segmentation models with user annotations.[18] Its results come from the studied image settings, not a universal validation for every pathology stain. The relevant lesson is that a pretrained model can be adapted and reviewed rather than accepted as a fixed measurement device. In a marker project, corrections should include the types of cells that change the scientific endpoint: tightly packed tumor cells, cells at gland boundaries and weakly stained candidates.

Annotation itself needs an uncertainty policy. Some objects are unambiguously tumor cells, some are clearly stromal or inflammatory cells, and others remain difficult to classify on the available section. A proposed annotation protocol should allow the third category instead of forcing every object into a confident positive or negative class. A reference set reviewed independently by more than one observer can document the frequency and location of disagreement. Consensus can improve consistency, but it should not erase the record of which cases were uncertain. That record is useful for evaluating model failures and planning targeted additional staining.

Copyright: ©2026 Mercan. 

## Quantification of measured immunostains

Immunohistochemistry (IHC) also offers defined molecular targets for digital analysis. DeepLIIF illustrates how deep learning can combine stain separation, segmentation and single-cell immunohistochemical scoring.[19] It is relevant as an enabling method, not as evidence that every cancer stem cell-associated antibody has already been validated with the platform. Transferring a method to a new stain changes the task: the expected subcellular localization, distribution of positive cells and tissue background may all differ. A useful evaluation would compare the automated output with an independently reviewed reference for that exact marker and specimen context.

The denominator must remain visible. Positive cells among all detected cells, positive tumor cells among all tumor cells and marker-positive area are different measurements. They cannot be combined under a single percentage without changing the biological question. Reporting the denominator, compartment and sampling rule also allows readers to identify whether an apparent increase reflects altered tumor-cell expression or a change in tissue composition.

## Learning from slides with limited cell level labels

Weak supervision addresses an annotation constraint that is particularly relevant to whole-slide images. Campanella and colleagues demonstrated large-scale computational pathology using slide-level supervision rather than exhaustive manual annotation of every region.[20] The enabling principle is important: a model can learn a slide-associated task while only some regions contain the informative pattern. Transferring this principle to stemness research requires caution. A slide-level molecular score gives information about the sampled tissue as a whole, while the model may assign high importance to regions that correlate with that score for several reasons.

The clustering-constrained attention multiple-instance learning method described by Lu and colleagues provides a further example of data-efficient learning from slide labels.[21] Its applications included tumor classification and metastasis detection, not validation of all CSC markers. For a stemness-related application, the aggregation model and the sampling strategy need to be described together. If only a subset of tiles is evaluated, rare marker-associated regions may be missed. If all tiles inherit a bulk label during a later analysis, the study risks giving local interpretations to a label that was never defined locally.

## Foundation models can improve representation without defining stemness

UNI was developed as a general-purpose pathology representation model and evaluated across multiple computational pathology tasks.[22] Prov-GigaPath was designed to learn from whole-slide data while retaining broader slide context.[23] These developments expand the available image representations. They do not remove the need to define a CSC-associated target, verify its reference assay or test performance in an independent cohort. A pretrained representation is an input to a study design, not evidence that the biological endpoint has already been validated.

For a small laboratory cohort, the key comparison is consequently task-specific. A foundation-model representation, a conventional image feature set and a simpler baseline can be compared under the same patient partitions and endpoint definition. Improvements should be judged on the independent test set and on errors that matter to the research question, not only on development-set discrimination. The encoder version and preprocessing should be retained so the result can be reproduced. If pretraining data overlap with an evaluation cohort, that possibility should be investigated and reported. Greater pretraining scale alone cannot establish independence of the final evaluation.

## What the application studies actually demonstrate

### Early morphological classification

Aichinger and colleagues evaluated 454 liver image regions from 43 patients, including 34 with cancer and nine healthy individuals.[24] Their conference report used serial-section immunostaining to motivate CSC-associated morphology and leave-one-patient-out evaluation of H&E classification. Color deconvolution plus a convolutional neural network achieved image-region accuracy of 0.936. Patient separation limits within-patient leakage, but the reported cancer-versus-healthy sampling and region-level endpoint do not demonstrate discrimination of functionally defined CSCs from other malignant cells. Neither external-cohort testing nor a functional cellular reference was established by this experiment. The result is feasibility evidence for the specified image classification setting.

### Morphometry associated with molecular stemness

Blocker and colleagues examined 49 patients with head and neck squamous cell carcinoma (HNSCC), relating automated nuclear measurements to matched transcriptomic scores.[25] Nuclear solidity, minimum diameter and circularity were associated with transcriptomic stemness after multiple-comparison correction. The molecular reference is stronger than an unverified visual stemness label, but it is a tissue-level construct and supplies no functional cell-by-cell truth. The image-to-stemness finding is an association in the analyzed cohort; the verified report does not establish external replication of that association. Neither the number of detected nuclei nor segmentation evaluation increases the independent patient sample beyond 49.

### Spatial phenotyping with a CD133-containing panel

Peng and colleagues studied 553 resected non-small cell lung cancer (NSCLC) cases of stage IA–IIIB at one institution, using multiplex immunofluorescence with a CD133-containing panel and StarDist-supported localization.[26] The prognostic analysis split these cases into 387 training and 166 testing patients. This is internal testing, despite the cohort size. Four primary cases underwent single-cell RNA sequencing; external transcriptomic corroboration does not constitute external validation of the image-based risk score. CD133 specificity and tumor-compartment annotation were acknowledged limitations. The work supports spatial phenotyping and prognostic associations, while functional CSC identity and causal cell communication remain unproven.

### Integrative pathomic signatures

Sun and colleagues combined single-cell and bulk transcriptomics with uveal melanoma whole-slide features to construct a nine-feature pathological image-derived prognostic score.[27] The source explicitly limits development and validation to a single TCGA-UVM cohort and calls for independent multicenter evaluation. The abstract-level source verified for this revision does not supply a sufficiently detailed image-analysis denominator or partition sizes, so these are not inferred

Copyright: ©2026 Mercan. 

from the nominal TCGA cohort. The supported result is an internally evaluated prognostic association. Computational drug-response or immune-response estimates remain hypotheses, and a prognostic map is not a cellular CSC map.

Molecular stemness itself is a constructed target. Malta and colleagues developed expression- and methylation-based indices of oncogenic dedifferentiation using a one-class machine learning approach.[28] Those indices are valuable research constructs, but predicting one from an image is not equivalent to measuring self-renewal in the corresponding cells. For an image model, the score definition, source material and level of aggregation should be stated. A bulk tissue label does not automatically provide a reliable annotation for every tile sampled from that tissue.

## Spatial architecture beyond a marker percentage

In this review, spatial mapping means retaining tissue coordinates while localizing measured or predicted phenotypes and quantifying their distribution within a defined compartment. Its scope includes cellular marker positions, cell-neighborhood composition, distances to cells or structures, regional enrichment, and image-linked molecular predictions at their validated resolution. It excludes treating an attention heatmap as a direct molecular measurement or assigning a functional CSC niche from proximity alone. Imaging mass cytometry[29] and histoCAT[30] provide enabling approaches for spatially resolved measurement and analysis; their use does not itself validate the biological identity of a CSC-associated population.

Spatial analysis should begin with a defined unit and comparison. A cell can be described by its own marker profile, its distance to a compartment boundary, or the composition of a neighborhood around it. Each description answers a different question. A nearest-neighbor calculation is sensitive to cell density, whereas a regional abundance measurement is sensitive to the boundary of the region. Comparing these outputs without their definitions can make studies appear contradictory even when they describe different aspects of the same tissue. Figure 2 uses a schematic example to show how equal marker frequency can coexist with different arrangements.

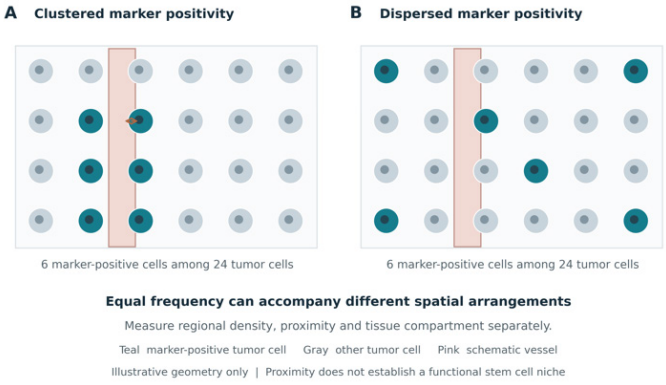


**Figure 2** Equal marker frequency can conceal different tissue arrangements.

Three levels of mapping should be reported explicitly. A cell-level map requires reliable segmentation, phenotype assignment and a tumor-cell denominator. A neighborhood map additionally requires a definition such as a distance radius, nearest-neighbor rule or graph connectivity. A regional or transcriptomic-spot map supports conclusions at that resolution unless cellular attribution is independently established. For each, the report should name the coordinate units, compartment mask, sampling window and treatment of boundaries. A proposed perivascular niche additionally needs independent vascular identification; a diagrammed vessel cannot establish a biological niche. These are proposed reporting specifications, not claims that all reviewed studies implemented them.

Original schematic contrasting clustered and dispersed marker positivity. Teal cells denote marker-positive tumor cells, gray cells other tumor cells, and the pink structure an illustrative vessel. Each panel contains six positive cells among 24 tumor cells (25%). The arrangement is conceptual; no patient coordinates, computed distance statistic or biological dataset is presented. In a real study, marker definition, tumor and vascular masks, distance units, neighborhood rule and sampling window would require specification. The schematic supports the distinction between abundance and arrangement; it does not demonstrate a functional niche, signaling direction or causal interaction.[5,26,29,30]

For a proposed neighborhood study, the null comparison should preserve relevant tissue constraints. Randomly mixing labels across an entire slide can create a misleading benchmark if tumor epithelium, necrosis and stroma occupy different regions. An analysis may instead evaluate relationships within specified compartments or preserve local structure when constructing a reference distribution. The rationale should be stated before interpreting enrichment or avoidance. A significant proximity pattern supports a spatial association under the chosen comparison; it does not identify the direction of signaling or prove that either cell population maintains the other.

Serial-section registration introduces another source of uncertainty. A cell seen on an H&E section may not appear in the same form on an adjacent immunostained section, and some cells may be present on only one section. Tissue distortion can further affect local alignment. Consequently, matched-region analyses and cell-by-cell correspondence should not be treated as equivalent. When the endpoint requires exact cellular co-expression, same-section multiplex measurement can answer a different question from the visual alignment of neighboring sections. The limitation belongs in both the methods and the interpretation of any training labels produced from that alignment.

## Linking image features to spatial molecular measurements

He and colleagues developed an approach for predicting spatial gene expression from breast tumor histology using paired spatial transcriptomic and image information.[31] This is an enabling example of linking morphology to local molecular measurements. It should not be cited as direct proof that H&E can determine every CSC-associated program. A spatial spot can contain multiple cells, and prediction at that spot resolution does not automatically assign the predicted signal to one cellular population. The granularity of the molecular reference therefore constrains the granularity of the conclusion.

A defensible extension to stemness research would first define the molecular program and tissue scale of interest. For example, a study might evaluate whether a region-level image representation predicts a predefined transcriptomic signature in independent tissue. The result should remain a region-level association unless cellular attribution is separately supported. The same data should not be used repeatedly to choose the signature, select the most favorable image features and then claim independent confirmation. Cross-modal agreement becomes stronger evidence when the molecular and imaging analyses are planned with genuinely separate evaluation material.

**Citation:** Mercan G. Digital pathology and artificial intelligence in cancer stem cell research: from quantification to spatial mapping. *J Stem Cell Res Ther.* 2026;11(1):120–130 DOI: 10.15406/jsrt.2026.11.00227

Copyright: ©2026 Mercan. 

## Virtual staining changes the status of the evidence

Pati and colleagues trained VirtualMultiplexer using an EMPaCT prostate cancer tissue microarray cohort of 210 patients, generating six virtual immunostains including CD44 from H&E.[32] Unpaired real IHC supplied the staining domain; evaluation included image fidelity, pathology review and downstream prediction. An independent EMPaCT subset supported internal testing, with additional external transfer and endpoint experiments. These broader evaluations deserve recognition, but they do not demonstrate cellwise CD44 truth or functional CSC detection in every transferred setting. Image realism and downstream performance address different questions from a paired, marker-specific assay validation. Virtual CD44 remains predicted expression, even where other model outputs transfer successfully.

The distinction matters most when a rare or spatially restricted population is the research target. Visual realism does not guarantee that the correct cell has been assigned the correct marker status. For a proposed cancer stem cell application, the review synthesis is to prioritize matched real-stain agreement, preservation of rare positive populations and errors in the tumor compartment over image appearance alone. Where sections are serial rather than identical, registration uncertainty also needs to be retained. Virtual staining may guide which regions to investigate, but confirmatory measurement is needed when the conclusion depends on actual antigen expression.

Virtual staining also changes the provenance of downstream data. If a generated CD44 image is used to calculate a spatial signature, both the signature and its uncertainty depend on the image-generation model. Training another predictor on that signature does not create an independent protein measurement. This matters when several apparently separate analyses all originate from the same H&E input. Agreement among them can be informative about model consistency while still leaving the original molecular claim untested. Reports should show which variables were measured experimentally and which were generated, transformed or inferred.

The proposed safeguard is to keep paired real measurements available for the claims that require them. A virtual image may guide region selection or support an exploratory association analysis. If it is used to assert disappearance of a rare marker-positive population after treatment, real staining and sampling controls are needed. Evaluation should include the relevant failure mode, such as missed rare positives or false positivity in a neighboring compartment. A realistic average image can conceal precisely the local error that would reverse a biological conclusion. This is why the reference standard should be chosen from the claim rather than from the convenience of the available data.

## Comparing reference standards and independence

The five applications cannot be ranked by a common accuracy statistic because their targets differ. The liver classifier evaluates region labels,[24] whereas the HNSCC study tests agreement between morphometric features and a bulk molecular construct.[25] The NSCLC investigation quantifies marker-associated spatial phenotypes and relates them to recurrence;[26] the uveal melanoma work produces a patient-level prognostic score.[27] VirtualMultiplexer evaluates generated stains and downstream predictions.[32] A large number of tiles, cells or cores can improve measurement coverage without creating more independent patients. Accordingly, Table 2 separates the biological sample from the image unit and identifies the reference and testing design for each application.

Independence also has several meanings. Patient-separated resampling estimates performance on unseen patients drawn from the development source; it does not by itself test changes in institution, staining or acquisition. An external transcriptomic dataset can corroborate a molecular relationship without testing an image algorithm. Conversely, external transfer of a virtual-staining pipeline does not validate every marker-specific cellular interpretation. Cohort selection, reference specificity and spatial resolution therefore remain separate appraisal dimensions. None of the five application reports, as reviewed here, establishes that using its image-derived CSC-associated output to choose treatment improves patient outcomes. These distinctions justify further research while limiting claims about functional identity and clinical use.

Table 2 Counts refer to patients unless otherwise specified; image regions, nuclei and tissue cores are not independent patients. “Not established” identifies a limitation of the evidence verified for this review and is not a formal risk-of-bias rating. NSCLC, non-small cell lung cancer; HNSCC, head and neck squamous cell carcinoma; UVM, uveal melanoma; IHC, immunohistochemistry; scRNA-seq, single-cell RNA sequencing.

**Table 1** Molecular targets and the limits of image-based interpretation

| Target or readout | What a defined assay can support | What needs additional evidence |
|---|---|---|
| CD133 or AC133 | Localization and abundance of the epitope recognized by the stated reagent | Equivalence to total CD133 or a functional CSC population[3,9] |
| ALDH1A1 or ALDH1A3 | Isoform-specific immunoreactivity in a processed specimen | Equivalence to viable-cell ALDH activity or tumor initiation[2,4] |
| Pan CD44 or a CD44 variant | The specified marker-associated phenotype and its tissue distribution | Isoform identity beyond reagent specificity or universal stem cell function[5,10] |
| LGR5-associated phenotype | Expression at the sampled time and location | Self-renewal, plasticity or metastatic contribution in that setting[6-8] |

**Citation:** Mercan G. Digital pathology and artificial intelligence in cancer stem cell research: from quantification to spatial mapping. *J Stem Cell Res Ther.* 2026;11(1):120–130 DOI: 10.15406/jsrt.2026.11.00227

Copyright: ©2026 Mercan. 

**Table 2** Cohorts reference standards and independent evaluation in selected applications

| Study and biological sample | Output and reference | Evaluation and independence | Interpretation boundary |
|---|---|---|---|
| Aichinger 2017[24]<br>Liver tissue<br>43 patients<br>454 image regions | H&E classification; serial-section immunostain-informed morphology. Cancer and healthy sampling. | Leave-one-patient-out internal evaluation. Region accuracy 0.936; no external cohort. | Feasibility of region classification; not a functional CSC count or malignant-cell-specific CSC assay. |
| Blocker 2022[25]<br>HNSCC<br>49 patients | Nuclear morphometry associated with matched tissue-level transcriptomic stemness. | Cohort association analysis; external replication of the stemness association not established in the verified source. | Bulk molecular correlation cannot identify the function of individual segmented cells. |
| Peng 2023[26]<br>NSCLC<br>553 patients at one institution | CD133-containing multiplex phenotypes, cell positions and recurrence; four primary scRNA-seq cases. | Internal prognostic split 387/166. External transcriptomic corroboration differs from external image-model testing. | Marker specificity and compartment masks limit identity claims; no externally validated risk score or demonstrated causal niche. |
| Sun 2026[27]<br>TCGA-UVM<br>Single source cohort | Nine-feature pathomic prognostic score linked to molecular stemness constructs. | Internal validation only. Image-analysis denominator and split sizes not resolved in the verified abstract. | Prognostic association and in silico treatment hypotheses; not measured cellular stemness or treatment benefit. |
| Pati 2024[32]<br>Prostate EMPaCT<br>210 patients; four cores per patient | Six virtual stains including CD44; unpaired real IHC, image fidelity, expert review and downstream endpoints. | Internal EMPaCT subset plus external transfer and endpoint studies; evaluation varies by task and marker. | External pipeline evidence is not universal cellwise CD44 or functional CSC validation. |

## Validation should follow the claim

### Keep patients and sites separate

The independent unit must be defined before images are divided into development and test sets. Multiple tiles, sections or blocks from one patient share biological and technical information. In the proposed workflow, all material from that patient is assigned to the same partition, and preprocessing and threshold choices are fitted using development data. An untouched test set should not become a second tuning set after disappointing results. Patient-level uncertainty estimates are preferable to treating thousands of related image tiles as thousands of independent patients.

Site effects deserve a separate test. Howard and colleagues demonstrated that deep learning can recover site-specific digital histology signatures and that these signatures can bias the apparent prediction of biological and clinical features.[33] The finding is enabling evidence about model validity rather than cancer stem cell-specific evidence. Its practical implication is that a model should be challenged on external specimens and acquisition conditions. Color normalization alone cannot be assumed to remove site information. Scanner, staining and specimen differences should be reported and examined rather than concealed by a pooled average.

### Match the reference standard to the output

A segmentation model needs object-level evaluation against reviewed boundaries. A marker-scoring model needs agreement with a defined, validated staining reference. A molecular prediction model needs comparison with independently measured molecular data. A claim about functional stemness needs appropriate biological experiments. These are proposed task-specific requirements, not a universal numerical performance threshold. The strongest reference for one task may be inadequate for another: pathologist agreement can establish reproducible interpretation of a stain without proving the function of the stained cells.

The CAP guideline on immunohistochemical assay validation remains relevant to the measured stain that supplies reference labels.[12] Digitization does not repair an inadequately characterized antibody or specimen protocol. Conversely, a well-performing stain does not establish that the image algorithm can separate tumor cells from background. Assay validation and computational evaluation should therefore be documented independently before they are combined into a biological claim.

### Use metrics that expose the important errors

For a rare positive population, overall accuracy can be high even when many relevant cells are missed. The proposed evaluation should retain sensitivity and positive predictive value at the intended level of analysis, together with the frequency and nature of false positives. For continuous marker scores, agreement and systematic bias are more informative than correlation alone. A model can correlate strongly with manual scores while consistently overestimating them. Probabilistic risk models also need calibration assessment, because ranking patients and estimating their risk are different properties.

Class imbalance deserves explicit analysis. Saito and Rehmsmeier showed why precision–recall evaluation can be more informative than an ROC display for imbalanced classification tasks.[34] In the present context, a rare marker-associated phenotype makes the proportion of true positives among predicted positives especially relevant. A study should retain the prevalence in the evaluation material and explain any enrichment of positive regions. Performance on a deliberately balanced set of selected image patches cannot be read as performance in an unselected clinical slide population. The operating threshold should be fixed before the final test.

Neural-network confidence also requires evaluation rather than interpretation at face value. Guo and colleagues studied calibration of modern neural networks and showed that good classification accuracy need not imply well-calibrated probabilities.[35] Their experiments were general machine-learning work, not validation of a CSC pathology assay. The applicable methodological point is that a value near one does not, by itself, make a cellular label certain. In a pathology study, calibration should be examined in the intended evaluation setting. Any recalibration procedure must be fitted without using the final test outcomes to choose the most favorable settings.

**Citation:** Mercan G. Digital pathology and artificial intelligence in cancer stem cell research: from quantification to spatial mapping. *J Stem Cell Res Ther.* 2026;11(1):120–130 DOI: 10.15406/jsrt.2026.11.00227

Spatial analyses require their own error analysis. Distances can change after segmentation, registration or tumor-mask errors. Stability should be examined under reasonable perturbations and alternative sampling choices. A proximity result should be accompanied by the cell definitions and tissue compartments that produced it. These recommendations are the review synthesis and should be tailored to the actual experiment rather than implemented as an unvalidated composite quality score.

## Interpretability needs its own evaluation

An attention heatmap can help locate image regions associated with a model output, but it is not a direct molecular map. Adebayo and colleagues demonstrated that some saliency methods can remain insensitive to changes in model parameters or training labels.[36] That finding concerns explanation methods in general machine learning and should be applied as a methodological caution rather than a pathology-specific performance claim. For CSC-associated research, visually plausible localization should be compared with independent tissue annotations or molecular measurements when those are the claimed explanation.

A useful explanation analysis should ask a falsifiable question. Does the highlighted region contain the independently measured marker? Does the explanation change when the relevant region is removed? Does it remain stable under a nuisance change that should not affect the biological endpoint? Such tests should be selected for the model and scientific task. Agreement with a pathologist's visual impression can support face validity, while randomization or perturbation analyses examine a different property. An explanation becomes more informative when its limitations are stated alongside its most persuasive example.

## Clinical utility requires a decision context

A prognostic association can be scientifically interesting without changing care. Decision curve analysis, introduced by Vickers and Elkin, evaluates prediction strategies in terms of net benefit across decision thresholds.[37] Its use requires a meaningful clinical decision and plausible consequences for false positives and false negatives. It should not be added to an exploratory marker paper merely as another performance plot. If the intended action is extra sampling, a confirmatory assay or a treatment choice, that action and its consequences should be defined first.

Treatment-response claims need particular restraint. A computational association with drug sensitivity or an inferred immunotherapy score is not equivalent to evidence that patients benefit from using the model to select treatment. External prediction evaluation is one step; a study of the model-guided decision is another. Before either, the measured or inferred tissue endpoint must be sufficiently stable to interpret. This review therefore treats clinical utility as a separate evidential question rather than the automatic destination of every successful image classifier. Research measurements can be valuable without being ready to guide individual patient care.

## Report a diagnostic study and a prediction study appropriately

STARD-AI addresses reporting of AI-centered diagnostic accuracy studies [38]. TRIPOD+AI addresses studies developing or evaluating clinical prediction models.[39] They are useful when the corresponding study purpose applies; neither is a substitute for study design or laboratory validation. An exploratory segmentation exercise should not claim clinical readiness simply because its manuscript follows a reporting checklist. For a marker-based prognostic model, the assay definition and the prediction-model description are both necessary. Table 3 translates these distinctions into a proposed planning framework.

**Table 3** Proposed validation plan organized by the intended output

| Intended output | Reference and evaluation | Claim that still needs more evidence |
|---|---|---|
| Cell or nuclear segmentation | Independent boundary annotations; object-level errors in difficult regions | Marker specificity and biological identity |
| Measured marker quantification | Validated stain; reviewed tumor masks; scoring agreement and bias | Functional self-renewal or tumor initiation |
| Virtual marker expression | Matched real staining; rare-positive and compartment-specific error analysis | Actual antigen expression in each predicted cell |
| Image-derived molecular stemness | Independent molecular measurements at the stated spatial scale | A functional CSC count |
| Prognostic or treatment-related model | Independent patient cohorts; discrimination, calibration and appropriate clinical evaluation | Improved outcomes or treatment benefit from using the model |
| Spatial phenotype or neighborhood map | Reviewed cell and compartment labels; declared coordinate scale and neighborhood rule; robustness to segmentation, registration and sampling | Causal cell interaction or a functional CSC niche |

Reporting guidance should also be distinguished from risk-of-bias assessment. PROBAST+AI provides an updated approach for assessing quality, risk of bias and applicability in prediction-model development and evaluation.[40] Its domains direct attention to participants and data sources, predictors, outcomes and analysis. Those concerns are relevant to an image-derived prognostic model, but a completed checklist does not establish that every relevant bias has been removed. In particular, a study may be transparently reported and still be too small, too selected or insufficiently independent to support its intended claim.

For a marker-associated prognosis study, REMARK remains relevant to the clarity of the assay, cohort and statistical reporting.[41] It complements rather than replaces prediction-model guidance when both purposes apply. A practical manuscript should identify its primary question and use the relevant reporting structure. Listing several acronyms in the methods without linking them to the actual

design adds little. The reader needs enough information to reconstruct what was measured, who contributed evaluation data, how choices were made and which inference depends on which source of evidence.

## A proposed framework for matching claims with evidence

The proposed contribution is a CSC-focused way to connect existing validation requirements across an image-to-claim workflow. It does not introduce a new validation principle, consensus standard or reporting checklist. Every major claim is described through four linked fields: the output, its reference standard, the evaluation population and the permitted level of interpretation. A marker abundance, a neighborhood score, a virtual stain, a molecular prediction and a clinical risk estimate require different evidence. The contribution is to make unsupported transitions among these outputs visible, especially when a marker-associated phenotype is described as a functional CSC. Figure 3 summarizes this author-proposed synthesis, which has not been prospectively evaluated.

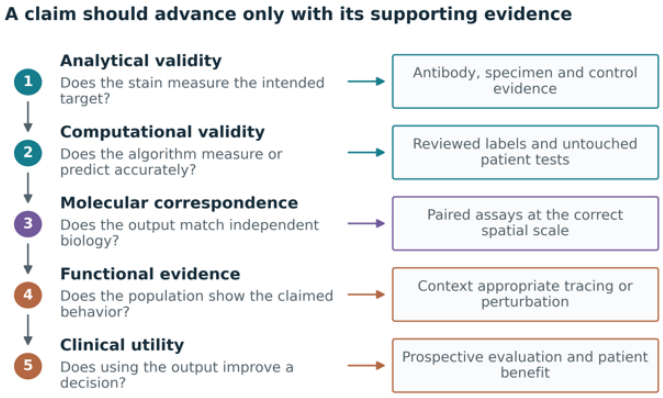


**Figure 3** Matching the scope of a claim to its supporting evidence.

The relationship to existing guidance is purpose-specific. CAP assay guidance addresses the validity of the measured IHC reference;[12] STARD-AI addresses reporting when the task is diagnostic accuracy;[38] TRIPOD+AI addresses clinical prediction-model reporting;[39] PROBAST+AI addresses quality, bias and applicability in prediction models;[40] and REMARK addresses tumor-marker prognostic study reporting.[41] The present framework connects the relevant guidance to a CSC-associated claim and additionally asks whether the reference supports marker measurement, molecular association or experimentally tested function. It neither substitutes for these documents nor applies every clinical checklist to an exploratory imaging task.

For example, a proposed model that maps CD133-associated cells near blood vessels would declare its output as the spatial distribution of a specified immunophenotype. Its reference would include reviewed real staining, tumor and vessel identification, and spatial annotations at a stated scale. Its evaluation population would be an independent set of patients with the required specimen characteristics. Its permitted interpretation would be a marker-associated spatial pattern, with uncertainty from segmentation and sampling. Calling the same map a self-renewing perivascular CSC niche would require additional context-appropriate functional experiments. This example demonstrates the framework's intended practical use without claiming that the illustrative design has been validated.

Original author-proposed synthesis informed by assay validation, functional experiments and reporting guidance.[6,8,12,37-41] Analytical validity, computational validity, molecular correspondence, functional evidence and clinical utility are distinct questions. The arrows indicate broader claims requiring additional evidence, not a mandatory sequence or a validated hierarchy. A research project may appropriately stop at a reproducible marker measurement. The framework has not been prospectively evaluated, consensus-endorsed or calibrated as a quality score.

For a measured phenotype, the central question is whether the assay and algorithm reproduce the stated observation. For an inferred molecular phenotype, the question becomes whether the output agrees with independent molecular measurements at the stated scale. For a functional claim, longitudinal or perturbational evidence appropriate to the biological system is needed. For a clinical claim, the effect of using the output must be evaluated in the relevant decision setting. These distinctions make an omission visible without dismissing the entire study: an exploratory association can remain a useful result when labeled accurately.

The framework can also guide a reviewer's interpretation of a striking figure. A colored spatial map should identify which colors represent observed staining and which represent predictions. A model-derived niche should specify how the neighborhood was defined and what independent evidence supports its biological role. A change after treatment should distinguish sampling variation from within-population change. The aim is not to add formalism for its own sake. It is to help readers locate the exact step at which a computational observation becomes a broader biological statement, and to judge whether that step is supported.

## A practical next study for a pathology laboratory

A tractable project would begin with one tumor type, one specified marker and one measurement question. For example, an investigator could evaluate automated measurement of CD133 immunoreactivity within pathologist-defined tumor regions. The clone, staining protocol, compartment and threshold rule would be fixed during development. The primary question would be whether the algorithm measures that immunophenotype reproducibly in independent specimens. Calling the endpoint immunophenotype avoids making an unsupported functional claim.

A second phase could relate those measurements to paired molecular data or experimental tumor models. The hypothesis and reference should be specified before testing: agreement with a transcriptomic signature, for example, answers a different question from tumor initiation. Sample size should be justified using the primary endpoint, prevalence and intended precision rather than the number of tiles that can be extracted. If the available cohort cannot support independent evaluation, the work should be presented as a pilot. No data or performance results from such a proposed study are reported here.

For teaching pathology laboratory techniques, the same design provides a useful critical exercise: students can trace how a tissue sample becomes a staining label, a cell mask and a reported score. The educational objective would be to identify the consequence of an error at each stage. This is a proposed application of the framework, not an evaluated educational intervention.

## Reproducibility and practical translation

A reproducible analysis requires more than the name of a software package. The report should retain the model checkpoint, preprocessing parameters, tissue masks, class definitions and rules used to aggregate cell-level measurements into a slide or patient

score. If manual corrections were made, the stage and purpose of those corrections should be stated. A script that reproduces a final plot is insufficient when key selections were made interactively and left undocumented. Conversely, a transparent record of a small, carefully defined experiment can be more useful than a large analysis whose endpoint changes during interpretation.

Independent testing should include material that resembles the intended research use. A dataset assembled from exceptionally clear positive and negative fields can be suitable for an early feasibility question but may exclude the very regions that cause problems in routine tissue. The report should describe exclusion criteria and the number of failed or unevaluable specimens. A future evaluation can then ask whether uncertainty handling, additional staining or human review makes the pipeline more reliable. Failure to produce an output is part of system performance and should not vanish from the denominator.

Implementation in a pathology laboratory also depends on workload and oversight. The time required to scan, annotate, review errors and repeat unsuccessful measurements should be recorded alongside computational speed. An AI method that saves counting time but requires substantial correction may still be useful, though its benefit differs from a fully automated workflow. A proposed pilot should therefore define the human role and retain a route for unresolved cases. These operational questions do not establish biological validity, but they determine whether a validated measurement can be used consistently outside the development exercise.

Shared materials can strengthen the field without requiring unrestricted release of identifiable clinical data. Representative de-identified failure examples, annotation manuals and versioned analysis settings can clarify how the result was obtained. When access is controlled, the conditions for independent evaluation should be explained. An institution can also evaluate whether a model remains stable after a documented change in staining or scanning. The general principle is traceability: readers and future investigators should be able to connect the reported result to the measurement choices that produced it.

## Limitations and research priorities

This purposive narrative review is vulnerable to selective retrieval, publication bias and uneven source access. It is not a systematic review, provides no estimate of retrieval completeness, and includes no formal study-level risk-of-bias assessment. The source-verification update strengthens specific comparisons but cannot establish that all relevant studies were identified. In particular, the verified uveal melanoma abstract does not resolve all image-cohort denominators and partition details. The application reports differ in tumor type, imaging unit, reference standard and evaluation design, so their metrics cannot be pooled or used to rank overall CSC-detection accuracy. Enabling technology is not direct CSC validation, internal testing is not external replication, and external molecular support is not necessarily independent evaluation of the image task. The proposed framework is an interpretive aid whose usability and reproducibility still require evaluation.

Priorities include paired image and molecular datasets, externally tested marker quantification and experiments that connect spatial phenotypes with function. Releasing representative failure cases, annotation rules and versioned analysis settings would make such work easier to scrutinize. The aim should be a reproducible link between a defined tissue observation and the biological question it can answer. Increasing model complexity without improving that link may add little to a laboratory's understanding of its samples.

## Conclusion

Digital pathology and AI can make specified CSC-associated phenotypes measurable through marker quantification, spatial phenotyping and molecular association studies. The selected applications provide heterogeneous evidence rather than a validated general method for detecting functional CSCs. Spatial mapping should identify whether its coordinates represent measured cells, inferred expression, neighborhood relationships or regional molecular predictions. A measured stain, a virtual stain and an image-derived stemness score must remain distinguishable. The proposed claim-to-evidence framework connects these distinctions with existing validation and reporting guidance, but its practical performance has not been tested. Independent patient and external-cohort evaluation, matched molecular references and context-appropriate functional experiments should be selected according to the claim. Clinical treatment selection requires separate evidence that using the output improves the relevant decision or outcome.

## Ethics and data availability

This review uses published literature and reports no newly collected human specimens, participant data or animal experiments. No new research dataset or AI model was generated or analyzed for this article.

## Acknowledgments

I dedicate this work to the cherished memory of my beloved brother, "Muzaffer MERCAN", who endured thirteen years of suffering as a result of a misdiagnosis and passed away on 10 March 2022.

## Conflicts of interest

The authors declare that there is no conflict of interest.

## References

1. Al–Hajj M, Wicha MS, Benito–Hernandez A, et al. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci USA*. 2003;100(7):3983–3988.
2. Ginestier C, Hur MH, Charafe–Jauffret E, et al. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell*. 2007;1(5):555–567.
3. Kemper K, Sprick MR, de Bree M, et al. The AC133 epitope, but not the CD133 protein, is lost upon cancer stem cell differentiation. *Cancer Res*. 2010;70(2):719–729.
4. Marcato P, Dean CA, Pan D, et al. Aldehyde dehydrogenase activity of breast cancer stem cells is primarily due to isoform ALDH1A3 and its expression is predictive of metastasis. *Stem Cells*. 2011;29(1):32–45.
5. Liu S, Cong Y, Wang D, et al. Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Reports*. 2014;2(1):78–91.
6. Schepers AG, Snippert HJ, Stange DE, et al. Lineage tracing reveals Lgr5+ stem cell activity in mouse intestinal adenomas. *Science*. 2012;337(6095):730–735.
7. Shimokawa M, Ohta Y, Nishikori S, et al. Visualization and targeting of LGR5+ human colon cancer stem cells. *Nature*. 2017;545(7653):187–192.

8. de Sousa e Melo F, Kurtova AV, Harnoss JM, et al. A distinct role for Lgr5+ stem cells in primary and metastatic colon cancer. *Nature*. 2017;543(7647):676–680.
9. Hermansen SK, Christensen KG, Jensen SS, et al. Inconsistent immunohistochemical expression patterns of four different CD133 antibody clones in glioblastoma. *J Histochem Cytochem*. 2011;59(4):391–407.
10. Boman BM, Viswanathan V, Facey COB, et al. The v8–10 variant isoform of CD44 is selectively expressed in the normal human colonic stem cell niche and frequently is overexpressed in colon carcinomas during tumor development. *Cancer Biol Ther*. 2023;24(1):2195363.
11. Kholodenko IV, Saidova AA, Potashnikova DM, et al. Contrasting impacts of targeted disruption of the cancer stem cell marker CD133 and its epigenetic regulator TRIM28 in colorectal cancer cells. *Int J Mol Sci*. 2025;26(22):10862.
12. Goldsmith JD, Troxell ML, Roy–Chowdhuri S, et al. Principles of analytic validation of immunohistochemical assays: guideline update. *Arch Pathol Lab Med*. 2024;148(6):e111–e153.
13. Moore HM, Kelly AB, Jewell SD, et al. Biospecimen reporting for improved study quality (BRISQ). *J Proteome Res*. 2011;10(8):3429–3438.
14. Yildiz–Aktas IZ, Dabbs DJ, Bhargava R. The effect of cold ischemic time on the immunohistochemical evaluation of estrogen receptor, progesterone receptor, and HER2 expression in invasive breast carcinoma. *Mod Pathol*. 2012;25(8):1098–1105.
15. Torlakovic EE, Francis G, Garratt J, et al. Standardization of negative controls in diagnostic immunohistochemistry: recommendations from the international ad hoc expert panel. *Appl Immunohistochem Mol Morphol*. 2014;22(4):241–252.
16. Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7:16878.
17. Graham S, Vu QD, Raza SEA, et al. Hover–Net: simultaneous segmentation and classification of nuclei in multi–tissue histology images. *Med Image Anal*. 2019;58:101563.
18. Pachitariu M, Stringer C. Cellpose 2.0: how to train your own model. *Nat Methods*. 2022;19(12):1634–1641.
19. Ghahremani P, Li Y, Kaufman A, et al. Deep learning–inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nat Mach Intell*. 2022;4(4):401–412.
20. Campanella G, Hanna MG, Geneslaw L, et al. Clinical–grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309.
21. Lu MY, Williamson DFK, Chen TY, et al. Data–efficient and weakly supervised computational pathology on whole–slide images. *Nat Biomed Eng*. 2021;5(6):555–570.
22. Chen RJ, Ding T, Lu MY, et al. Towards a general–purpose foundation model for computational pathology. *Nat Med*. 2024;30(3):850–862.
23. Xu H, Usuyama N, Bagga J, et al. A whole–slide foundation model for digital pathology from real–world data. *Nature*. 2024;630(8015):181–188.
24. Aichinger W, Krappe S, Cetin AE, et al. Automated cancer stem cell recognition in H and E stained tissue using convolutional neural networks and color deconvolution. In: Gurcan MN, Tomaszewski JE, eds. Medical Imaging 2017 Digital Pathology. *Proc SPIE*. 2017;10140:101400N.
25. Blocker SJ, Cook J, Everitt JI, et al. Automated nuclear segmentation in head and neck squamous cell carcinoma pathology reveals relationships between cytometric features and ESTIMATE stromal and immune scores. Am J Pathol. 2022;192(9):1305–1320.
26. Peng H, Wu X, Liu S, et al. Multiplex immunofluorescence and single–cell transcriptomic profiling reveal the spatial cell interaction networks in the non–small cell lung cancer microenvironment. *Clin Transl Med*. 2023;13(1):e1155.
27. Sun Y, Wu J, Yuan Y, et al. Multi–omics integrative analysis of stemness–associated pathological signatures to guide prognosis and therapeutic strategies in uveal melanoma. *Int J Surg*. 2026;112(5):12044–12060.
28. Malta TM, Sokolov A, Gentles AJ, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*. 2018;173(2):338–354.e15.
29. Giesen C, Wang HAO, Schapiro D, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods*. 2014;11(4):417–422.
30. Schapiro D, Jackson HW, Raghuraman S, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat Methods*. 2017;14(9):873–876.
31. He B, Bergenstrahle L, Stenbeck L, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng*. 2020;4(8):827–834.
32. Pati P, Karkampouna S, Bonollo F, et al. Accelerating histopathology workflows with generative AI–based virtually multiplexed tumour profiling. *Nat Mach Intell*. 2024;6:1077–1093.
33. Howard FM, Dolezal J, Kochanny S, et al. The impact of site–specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12:4423.
34. Saito T, Rehmsmeier M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
35. Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. *Proc Mach Learn Res*. 2017;70:1321–1330.
36. Adebayo J, Gilmer J, Muelly J, et al. Sanity checks for saliency maps. *Adv Neural Inf Process Syst*. 2018;31.
37. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
38. Sounderajah V, Guni A, Liu X, et al. The STARD–AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283–3289.
39. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
40. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505.
41. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst*. 2005;97(16):1180–1184.