arXiv:2608.23803v1 [cs.CV] 24 Aug 2026

# LUCAID: Agentic Multimodal AI for Lung Cancer Precision Pathology

Marie-Lisa Eich[1,2,#], Kai Standvoss[3,#], Timo Milbich[3,#], Alexander Möllers[3,6,10,#], Miriam Hägele[3], Philipp Anders[4], Lars Tharun[4], Hanna Kontradiuk[3], Sebastian Kons[3], Nader Aldoj[3], Recepcan Adigüzel[3], Adam Narai[3], Lukas Hönig[3], Jonathan Striebel[3], Binru Yang[1], Mihnea P. Dragomir[1,13], Marvin Sextro[3,6,10], Philipp Keyl[1,10,11], Philipp Jurmeister[11], Rosemarie Krupar[3], Evelyn Ramberger[3], James Wells[3], Julika Ribbat-Idel[3], Andreas Kunft[3], Hussam Shuaib[5], Christian Grohé[5], Reinhard Büttner[15], David Horst[1], Klaus-Robert Müller[6,7,8,9,10], Lukas Ruff[3], Maximilian Alber[3,+], Frederick Klauschen[1,3,10,11,12,14,+], and Simon Schallenberg[1,13,+]

[1]Institute of Pathology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Germany

[2]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, BIH Biomedical Innovation Academy, BIH Charité Digital Clinician Scientist Program, Charitéplatz 1, 10117 Berlin, Germany

[3]Aignostics GmbH, Berlin, Germany

[4]MVZ HPH Institut für Pathologie und Hämatopathologie GmbH, Hamburg, Germany

[5]Evangelische Lungenklinik Berlin-Buch, Berlin, Germany

[6]Machine Learning Group, Technical University of Berlin, Berlin, Germany

[7]Department of Mathematics and Computer Science, Technical University of Berlin, Germany

[8]Department of Artificial Intelligence, Korea University, Seoul 136-713, South Korea

[9]MPI for Informatics, Saarbrücken, Germany

[10]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

[11]Institute of Pathology, Ludwig-Maximilians-University, Munich, Germany

[12]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Munich Partner Site, Heidelberg, Germany

[13]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Berlin Partner Site, Heidelberg, Germany

[14]Bavarian Cancer Research Center (BZKF), Munich Partner Site, Munich, Germany

[15]Institute of Pathology, University Hospital Cologne, Cologne, Germany

**Abstract**

Lung cancer tissue diagnostics is complex, as therapy decisions in precision oncology rely on the integration of histomorphological, immunohistochemical, and molecular features. Yet pathological assessment remains largely visual and semi-quantitative and shows interobserver variability, while existing artificial intelligence (AI) tools cover only selected tasks, rarely reach generalizable expert-level performance, and lack prospective clinical validation. To address these challenges, we developed and clinically validated LUCAID, an agentic AI system for precision lung cancer pathology. An integrative agent couples diagnostic reasoning with nine modules that cover the full routine workflow, from quality control, tumor detection and segmentation, histological subtyping, tumor microenvironment profiling, tumor cellularity quantification, and predictive biomarker scoring (PD-L1, MET, TROP-2) to automated structured report generation. LUCAID enables users to interactively query the module outputs and generate reports that contextualize the results. Against large-scale expert ground-truth annotations, the analysis modules achieved F1 scores of 0.82–0.95. In prospective clinical validation, LUCAID reached 93.0% concordance with an expert-panel adjudicated reference standard across clinically actionable decisions, compared with 68.3–81.1% for five experienced thoracic pathologists.

# 1 Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide, accounting for approximately 1.8 million deaths annually [1]. Recent advances in targeted therapies, immune checkpoint inhibitors, and antibody-drug conjugates (ADCs) have substantially expanded treatment options for patients with actionable biomarkers [2–8]. Consequently, lung cancer pathology has evolved from a predominantly morphological discipline into a multimodal field that increasingly depends on evaluating and integrating histopathological, immunohistochemical, and molecular information. Therapeutic decisions now require several distinct assessments: reliable identification of diagnostically evaluable tumor tissue, precise histological subtyping, quantitative assessment of subcellular biomarker expression, and integration of these measurements with molecular profiling results [9–15].

However, the rigorous and standardized quantification of these histopathological features remains challenging in the daily diagnostic workflow [16–18]. Furthermore, novel candidate biomarkers within the tumor microenvironment (TME), such as the density and spatial distribution of tumor-infiltrating lymphocytes [19, 20], are continuously emerging but are impractical to evaluate manually at scale. This complexity grows further with the immunohistochemical biomarkers required to guide therapeutic decisions, which are increasing in number and in the complexity of their assessment. The list of candidates is expanding rapidly beyond established targets such as programmed cell death-ligand 1 (PD-L1) and the ADC targets trophoblast cell-surface antigen 2 (TROP-2) and hepatocyte growth factor receptor (MET) [12, 21]. At the same time, even single-marker evaluation is demanding, as many markers are subject to multiple scoring systems, each requiring assessment of positivity in different cell populations and compartments [22–24]. This makes manual quantitative evaluation increasingly laborious and prone to interobserver variability [25–29]. Spatial relationships such as cell–cell distances and cellular neighborhoods are also largely inaccessible to routine visual evaluation but may provide an additional layer of biological characterization [30, 31].

Artificial intelligence (AI) can address these challenges, as pattern recognition algorithms can quantify even fine-grained features at scale. Recent progress has been driven by vision foundation models trained on large, heterogeneous histopathology datasets spanning diverse tissue types and staining modalities [32–36]. Fine-tuned for specific applications, these models have been shown to achieve competitive performance on selected pathology tasks [37, 38]. Yet two limitations constrain their clinical impact. First, many downstream applications do not reach pathologist-level performance, and the few that approach it have rarely been prospectively clinically validated. Second, they are built for isolated tasks and do not integrate multimodal information across the multi-step diagnostic workflow. Agentic systems have the potential to address the latter, but their use in pathology remains early: current systems cover only parts of the diagnostic workflow, rely on vision-language models for qualitative image interpretation rather than validated quantitative measurements, and lack clinical validation [39–42].

To address these limitations, we developed LUCAID (Lung Cancer Agent for Integrative Diagnostics), an agent-driven multimodal AI system for lung cancer pathology that couples interactive diagnostic reasoning with a suite of extensively validated quantitative analysis modules. LUCAID's modules are based on the Atlas family of histopathology models and here we extend them towards comprehensive lung cancer precision pathology [36, 38, 43]. They cover tissue quality control, tumor detection and subtyping, TME characterization, tumor cellularity assessment, immunohistochemical (IHC) phenotyping, and PD-L1 and ADC-target expression scoring. For each case, the agent selects and orchestrates the relevant diagnostic modules, interprets their outputs, and integrates them into a structured pathology report, with all quantitative measurements generated by the underlying analysis modules. Individual

[#] Contributed equally to this work
[+] Contributed equally to this work

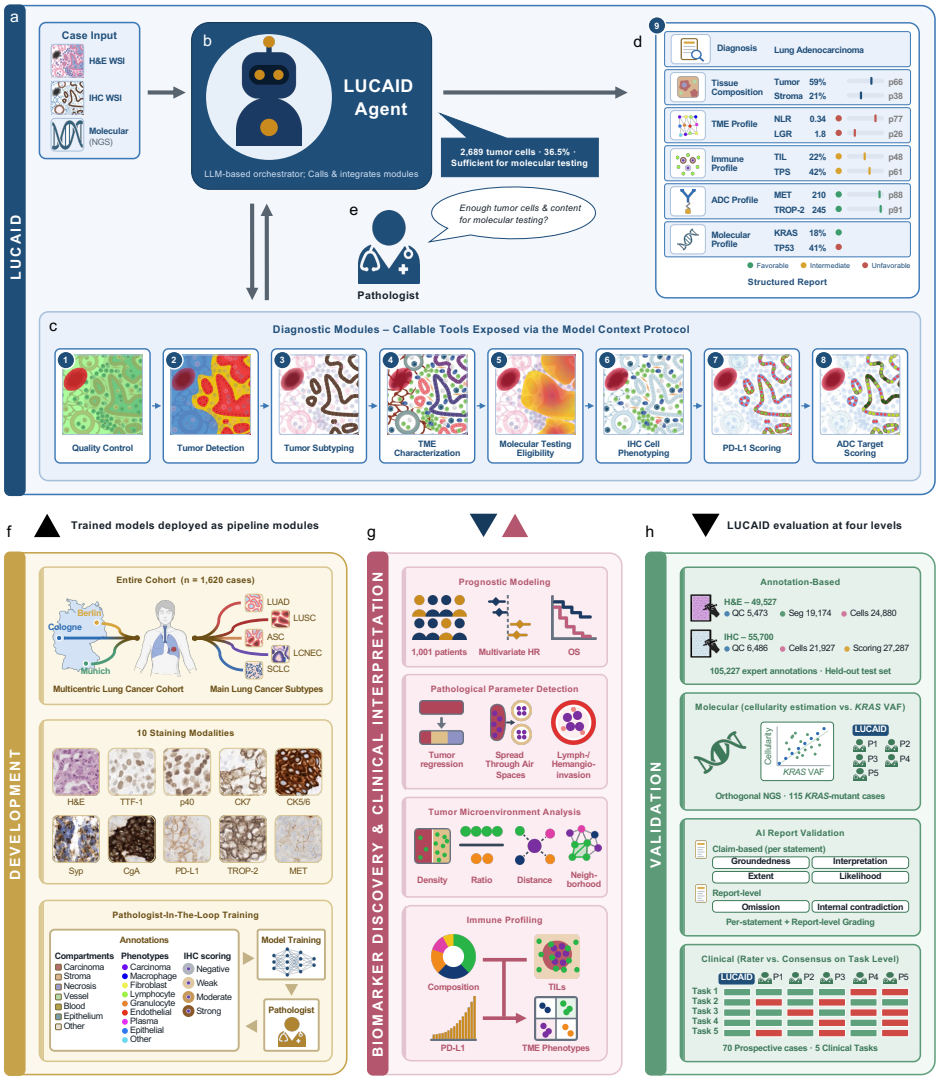



Figure 1: **LUCAID system overview, development, biomarker discovery and validation.**
**a**, Case input: digitized H&E and immunohistochemistry (IHC) whole-slide images (WSIs) from a routine lung cancer case, together with molecular profiling results, enter the system. **b**, A large language model (LLM)-based agent coordinates the analysis by calling diagnostic modules as needed and integrating their validated quantitative results. **c**, Analysis modules (left to right): (1) quality control (QC); (2) tumor detection via tissue segmentation; (3) tumor subtyping; (4) tumor microenvironment (TME) characterization; (5) tumor cellularity assessment; (6) IHC cell phenotyping; (7) PD-L1 scoring; and (8) antibody–drug conjugate (ADC) target expression scoring for TROP-2 and MET. **d**, The agent integrates the module results and generates a structured report comprising six sections: diagnosis, tissue composition, TME profile, immune profile, ADC profile and molecular profile. **e**, Throughout the analysis, the pathologist can pose case-specific clinical queries in natural language and receive quantitative, reproducible answers based on the module outputs. **f**, Development: a multicentric cohort of 1,620 lung cancer cases encompassing all major histological subtypes included digitized H&E and IHC WSIs spanning ten staining modalities. Models were trained using a pathologist-in-the-loop approach with H&E and IHC annotations. **g**, Biomarker discovery and clinical interpretation: LUCAID was applied to a discovery cohort of 1,001 patients to quantify established pathological features, investigate candidate biomarkers, characterize the tumor and immune microenvironment, and contextualize individual patient findings using cohort-level reference distributions. **h**, Validation at four levels: annotation-based benchmarking using 105,227 expert annotations across H&E and IHC; AI report validation using claim- and report-level assessment of groundedness, interpretation, extent, likelihood, omission and internal contradiction; molecular validation against orthogonal next-generation sequencing data from 115 *KRAS*-mutant cases; and prospective multicentric clinical validation in 70 cases across five pathologists and five clinically actionable tasks.
*Abbreviations:* ADC, antibody–drug conjugate; ASC, adenosquamous carcinoma; CgA, chromogranin A; CK5/6, cytokeratin 5/6; CK7, cytokeratin 7; H&E, hematoxylin and eosin; IHC, immunohistochemical; LCNEC, large cell neuroendocrine carcinoma; LLM, large language model; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MET, hepatocyte growth factor receptor; PD-L1, programmed death-ligand 1; QC, quality control; SCLC, small cell lung cancer; SYP, synaptophysin; TME, tumor microenvironment; TROP-2, trophoblast cell-surface antigen 2; TTF-1, thyroid transcription factor 1; WSI, whole-slide image.

components are extensively evaluated against expert annotations, while the integrated system is further assessed prospectively in a real-world clinical setting. Importantly, this validated modularity differentiates LUCAID from current vision-language models approaches that remain unreliable on fine-grained perception and counting task that quantitative pathology depends on [44–47]. Figure 1 provides an overview of LUCAID. Designed to operate in all major histological subtypes of lung cancer, LUCAID brings reproducible, quantitative AI to the full diagnostic workflow.

Using 1,620 lung cancer cases from the Institutes of Pathology at Charité – University Medical Center Berlin, University Hospital Cologne, and Ludwig Maximilian University Munich, together with an independent prospective multicentric validation cohort from the National Network Genomic Medicine Lung Cancer (nNGM), we show that LUCAID's modules achieve expert-level performance against large-scale ground-truth annotations. In prospective validation across clinically actionable diagnostic tasks, LUCAID achieved 93.0% concordance with an expert-panel adjudicated reference standard, compared with 68.3–81.1% for individual pathologists. Building on this foundation, we demonstrate how LUCAID integrates these modules into end-to-end case analysis, pointing toward a new generation of integrated computational pathology systems.

## 2 Results

LUCAID supports routine pathology workflows in lung cancer diagnosis by agentic orchestration of nine AI modules for comprehensive analysis of hematoxylin & esoin (H&E)- and IHC-stained WSIs: (1) *tissue quality control* to identify analyzable tissue and to discard artifacts on both H&E- and IHC-stained whole-slide images (WSIs); (2) *tissue segmentation* for detecting tumor regions and delineation of tissue compartments; (3) *tumor subtyping* from expressions of specific IHC markers; (4) *TME profiling* by classifying individual cells on H&E to characterize the tumor microenvironment; (5) *tumor cellularity assessment* for tumor content quantification to guide molecular testing; (6) *IHC cell phenotyping* by classifying cell identities IHC-stained WSIs; (7, 8) *biomarker expression scoring* to quantify the expression of the prognostic biomarker PD-L1 (7) and the ADC targets MET and TROP-2 (8) at cell and subcellular resolution; and (9) structured report generation, which integrates the quantitative results of the applicable diagnostic modules into a case-level pathology report. Each module can be called on demand during interactive pathologist–LUCAID interaction or orchestrated end-to-end to cover the diagnostic workflow from H&E-based tissue profiling to immunohistochemical assessment and molecular testing.

To extensively assess LUCAID's capabilities, we validated it at four different levels (Figure 1). First, each of the eight analysis modules was evaluated on dedicated hold-out test sets against expert ground-truth annotations. The underlying multicentric dataset comprised 1,620 lung cancer cases from the Institutes of Pathology at Charité – University Medical Center Berlin, University Hospital Cologne, and Ludwig Maximilian University Munich and included H&E-stained slides as well as nine different IHC markers; task-specific subsets were used for module-level testing. The dataset includes lung adenocarcinomas (LUADs), lung squamous cell carcinomas (LUSCs), adenosquamous carcinomas (ASCs), large cell neuroendocrine carcinomas (LCNECs), and small cell lung carcinomas (SCLCs). Second, to assess biological concordance, outputs of selected modules were benchmarked against orthogonal molecular reference measurements. Third, we evaluated the system's text- and report-generation capabilities by assessing whether the generated reports remain grounded in the validated module measurements. Fourth, we prospectively evaluated end-to-end execution of LUCAID in an independent lung cancer cohort from the National Network Genomic Medicine (nNGM) and benchmarked its clinically actionable diagnostic predictions against five experienced thoracic pathologists.

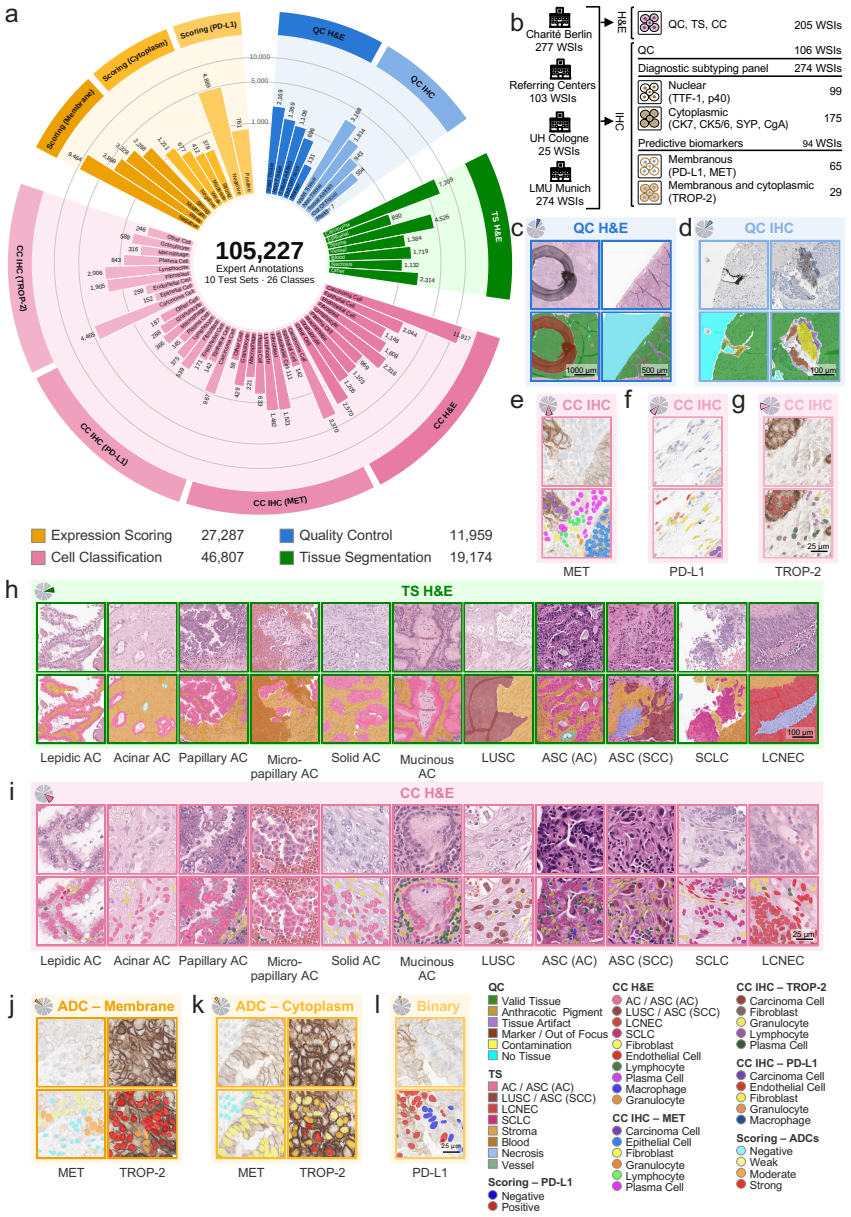



Figure 2: **Expert annotation test sets for LUCAID module validation.**
**a**, Distribution of 105,227 expert annotations across ten independent hold-out test sets, comprising 26 prediction classes in four algorithm categories: quality control, tissue segmentation, cell classification and expression scoring. Annotation counts are shown for each prediction class. **b**, Overview of the multicentric whole-slide image test sets, showing contributing institutions and their use for H&E-based analysis, diagnostic subtyping and predictive biomarker evaluation. **c**–**l**, Histopathological images are shown above with the corresponding expert annotations below. **c**,**d**, H&E and IHC examples for quality control. **e**–**g**, IHC cell-classification examples for MET, PD-L1 and TROP-2, respectively. **h**, H&E tissue-segmentation examples across major lung carcinoma subtypes and six adenocarcinoma subtypes. **i**, Corresponding H&E cell-classification examples. **j**,**k**, Subcellular expression-scoring examples for membranous and cytoplasmic MET and TROP-2 expression. **l**, Binary PD-L1 expression-scoring example.
*Abbreviations:* AC, adenocarcinoma; ADC, antibody–drug conjugate; ASC, adenosquamous carcinoma; CC, cell classification; CgA, chromogranin A; H&E, hematoxylin and eosin; IHC, immunohistochemistry; LCNEC, large cell neuroendocrine carcinoma; LMU, Ludwig Maximilian University; LUSC, lung squamous cell carcinoma; MET, hepatocyte growth factor receptor; PD-L1, programmed death-ligand 1; QC, quality control; SCC, squamous cell carcinoma; SCLC, small cell lung cancer; SYP, synaptophysin; TS, tissue segmentation; TROP-2, trophoblast cell-surface antigen 2; TTF-1, thyroid transcription factor 1; WSI, whole-slide image.

## 2.1 Module Predictions Recover Expert Ground-Truth Annotations

The modules are introduced and validated in sequence of a routine diagnostic process. For tumor detection and cell phenotyping, we characterize performance on lung in detail, resolved across all five major subtypes – LUAD, LUSC, ASC, LCNEC, and SCLC – spanning the non-small-cell, neuroendocrine, and small-cell entities. Figure 2 summarizes annotation counts across the independent hold-out test sets, the multicentric WSI cohorts contributing to each diagnostic task, and representative expert annotations across quality control, tissue segmentation, cell classification and expression scoring.

### 2.1.1 Module 1: Quality Control

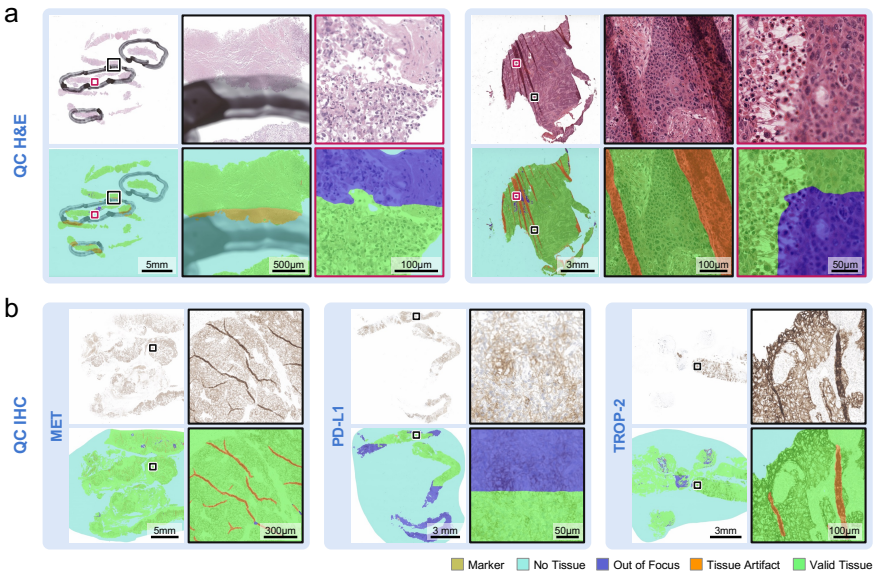
The quality control (QC) module automatically identifies tissue regions and common artifacts in both H&E- and IHC-stained WSIs, distinguishing valid tissue, out-of-focus regions, tissue artifacts, markers and non-tissue background. Representative applications illustrate robust artifact detection while preserving diagnostically relevant tissue regions (Figure 3a,b). We evaluated it on a dataset comprising 5,473 annotations from 132 H&E WSIs and 6,486 annotations from 106 IHC WSIs. Classification performance was high across QC classes, with overall F1 scores of 0.91 for H&E and 0.88 for IHC slides; valid tissue and non-tissue background each reached F1 scores of 0.99 in both staining modalities (Figure 3c). By maintaining robust performance across both staining modalities, LUCAID extends QC beyond previous H&E-focused tools [48–51], and provides a unified preprocessing step for all downstream analyses.

### 2.1.2 Module 2: Tissue Segmentation for Tumor Detection and Compartmentalization

The tissue segmentation module automatically detects tumor regions, delineates distinct tumor compartments, and classifies different tissue types within H&E-stained lung tissue (Figure 3d; metastatic samples shown in Supplementary Figure 2a), across all major lung cancer subtypes (LUAD, LUSC, ASC, SCLC, and LCNEC). It distinguishes seven categories: carcinoma, epithelium, stroma, necrosis, blood, vessel, and other tissue components commonly found in lung specimens, including alveolar lung parenchyma, cartilage, smooth muscle, nerves, and secretions such as mucus. Quantitative evaluation on a test dataset of 19,174 annotations from 205 WSIs, curated to span all five subtypes (Figure 2a+h), showed strong performance across all tissue categories. The model achieved an average F1 score of 0.93 (detailed class-wise results are provided in Figure 3e). Notably, carcinoma detection reached an F1 score of 0.98 across all carcinoma subtypes (range 0.97–0.99) and showed a comparable F1 score of 0.99 when evaluated on metastatic samples (range 0.97–0.99; Supplementary Figure 2b).This is important, as accurately identifying tumor regions across diverse lung cancer subtypes is a central diagnostic task.

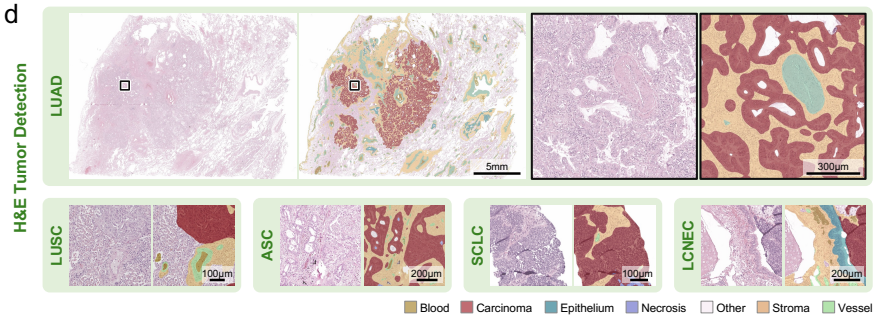
### 2.1.3 Module 3: Tumor Subtyping via Cell Phenotyping

To enable cell-level analysis on IHC-stained slides, we first trained a cell phenotyping model capable of detecting and classifying individual cells across nine prediction classes in IHC WSIs (see LUCAID module 6: cell phenotyping on IHC for further details (Figure 4)). Subsequently, an expression scoring model was trained to determine marker-specific expression status (positive vs. negative) exclusively in cells classified as tumor cells by the IHC cell phenotyping model. Cytoplasmic expression was evaluated for CK5/6, CK7, CgA, and SYP, whereas nuclear expression was assessed for TTF-1 and p40. Cases with >10% marker-positive tumor cells were classified as positive for the respective marker in accordance with WHO classification criteria for lung tumors (Figure 5a)[52]. Exemplary cases are shown in Supplementary Figure 1. Integration of tumor cell-specific expression profiles enabled




c

| Prediction Class | Out of Focus | Tissue Artifact | Marker | Valid Tissue | No Tissue | Average |
|---|---|---|---|---|---|---|
| F1 H&E | 0.70 | 0.92 | 0.96 | 0.99 | 1.00 | **0.91** |
| F1 IHC | 0.73 | 0.84 | 0.88 | 0.99 | 0.98 | **0.88** |




e

| Prediction Class | Carcinoma | Epithelium | Stroma | Necrosis | Blood | Vessel | Other | Average |
|---|---|---|---|---|---|---|---|---|
| F1 Overall | 0.98 | 0.86 | 0.95 | 0.92 | 0.95 | 0.89 | 0.95 | **0.93** |
| F1 LUAD | 0.98 | 0.92 | 0.93 | 0.96 | 0.97 | 0.92 | 0.96 | **0.95** |
| F1 LUSC | 0.97 | 0.77 | 0.94 | 0.89 | 0.91 | 0.85 | 0.84 | **0.88** |
| F1 ASC | 0.98 | 0.70 | 0.96 | 0.97 | 0.93 | 0.83 | 0.94 | **0.90** |
| F1 SCLC | 0.99 | 0.98 | 0.96 | 0.96 | 0.99 | 0.83 | 0.97 | **0.95** |
| F1 LCNEC | 0.99 | 0.98 | 0.95 | 0.94 | 0.67 | 0.94 | 0.94 | **0.92** |

Figure 3: **Quality control and tumor detection across lung cancer subtypes.**
**a**, H&E quality-control examples shown as histopathological images above and corresponding model-derived overlays below. Two biopsy specimens are shown, each with an overview image followed by two higher-magnification views. **b**, IHC quality-control examples for MET, PD-L1 and TROP-2, each shown as an overview image and a higher-magnification view, with histopathological images above and corresponding model-derived overlays below. The QC model distinguishes five classes: out of focus, tissue artifact, marker, valid tissue and no tissue. **c**, F1 scores for each QC class on H&E- and IHC-stained WSIs. **d**, H&E tumor-detection examples across five major lung carcinoma subtypes, with histopathological images shown on the left and corresponding model-derived tissue-segmentation overlays on the right. For LUAD, a resection specimen is shown with a higher-magnification view; the lower row shows LUSC, ASC, SCLC and LCNEC from left to right. The segmentation model distinguishes seven tissue compartments: carcinoma, epithelium, stroma, necrosis, blood, vessel and other. **e**, F1 scores for each tissue compartment overall and stratified by histological subtype.
*Abbreviations:* ASC, adenosquamous carcinoma; H&E, hematoxylin and eosin; IHC, immunohistochemistry; LCNEC, large cell neuroendocrine carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MET, hepatocyte growth factor receptor; PD-L1, programmed death-ligand 1; SCLC, small cell lung cancer; TROP-2, trophoblast cell-surface antigen 2.

automated tumor classification according to established diagnostic criteria, such as LUAD diagnosis based on TTF-1 and CK7 positivity combined with absence of p40 and CK5/6. For lineage markers

distinguishing LUAD from LUSC, LUCAID achieved accuracies of 93.3% for TTF-1 and 97.6% for CK7, as well as 98.1% for p40 and 100% for CK5/6 using pathologists' case level scoring as ground truth (Figure 5b). For neuroendocrine markers, accuracies reached 95.5% for SYP and 92.1% for CgA, resulting in consistently high classification performance across all diagnostic lineage markers (Figure 5b).

#### 2.1.4 Module 4: Cell Phenotyping for TME Profiling

To enable detailed analysis of the cellular TME, the H&E cell phenotyping module automatically detects and classifies individual cells within H&E-stained WSIs. The model distinguishes carcinoma cells from major inflammatory cell populations, including lymphocytes, macrophages, granulocytes, and plasma cells, as well as other non-neoplastic cell types such as epithelial cells, endothelial cells, and fibroblasts. Quantitative evaluation was carried out on a test dataset of 24,880 annotations across 205 WSIs, curated to span all five major lung cancer subtypes (LUAD, LUSC, ASC, LCNEC, and SCLC; Figure 2a+d). Representative cell classifications across all five major lung cancer subtypes are shown in Figure 4a, with metastatic samples provided in Supplementary Figure 2c. Quantitative evaluation demonstrated strong performance across all cell classes, with an average F1 score of 0.95 (range across subtype: 0.90–0.98) and average class-specific F1-scores ranging from 0.68 to 1.00 (Figure 4b). Notably, carcinoma cell identification achieved an F1 score of 0.95 in primary tumor and an F1 score of 0.99 in metastatic tumors (Supplementary Figure 2d), providing the basis for precise tumor cellularity assessment to guide molecular profiling (see section: LUCAID Module 5: Tumor Cellularity Assessment for Molecular Profiling).

#### 2.1.5 Module 5: Tumor Cellularity Assessment for Molecular Profiling

The LUCAID tumor cellularity assessment module quantifies quantifies tumor content in H&E-stained WSIs to guide molecular testing, using both cell count-based and nuclear area-based approaches.

To further capture spatial heterogeneity in tumor cellularity, we developed a tumor cell clustering-based visualization framework that generates spatial maps of local tumor content across WSIs. The resulting heatmaps provide an intuitive representation of regional variation in tumor cellularity, with a continuous color gradient ranging from low (yellow) to high (red) tumor content (Figure 5e and Supplementary Figure 3c). The representative tumor illustrates pronounced intratumoral heterogeneity that is readily apparent in the generated heatmap but may be difficult to appreciate by conventional H&E assessment alone. By visualizing regional variation in predicted tumor content, the heatmaps may support more standardized selection of tissue areas for molecular profiling and reduce variability associated with visual estimation.

#### 2.1.6 Module 6: IHC Cell Phenotyping for Biomarker Expression Scoring

The LUCAID IHC cell phenotyping module extends the H&E-based cell phenotyping of module 4 to IHC-stained WSIs, classifying individual cells into nine phenotypes. This allows biomarker expression to be assigned to specific cell populations rather than evaluated at the tissue level alone. Representative examples of expert test annotations are provided in Figure 2e–g. We evaluated the module on the three major therapeutic biomarkers in lung cancer: PD-L1 and the ADC targets TROP-2 and MET (n = 32 TROP-2; n = 35 PD-L1; n = 30 MET). The test set comprised 3,180 ground-truth annotations for PD-L1, 10,780 for TROP-2, and 7,967 for MET (Figure 2a). Annotation density reflected the increasing complexity of the respective scoring systems, from two categories for PD-L1 to four intensity levels for MET and four levels across both membranous and cytoplasmic compartments for TROP-2, with

additional annotations used to capture the resulting variation in staining and expression patterns in the hold-out sets. Representative classifications across the three stainings illustrate robust cell detection and phenotyping despite differences in tumor architecture, staining intensity, and subcellular expression patterns (Figure 4c). The model achieved F1 scores of 0.86 for TROP-2, 0.86 for MET, and 0.90 for PD-L1 across all cells (Figure 4d). Notably, carcinoma-cell identification reached F1 scores of 0.98 for TROP-2, 0.97 for MET, and 0.95 for PD-L1, supporting reliable assignment of biomarker expression to the tumor-cell compartment in which these markers are clinically assessed. Together, these results support cell-level biomarker quantification across distinct therapeutic IHC stainings.

#### 2.1.7 Modules 7 and 8: Biomarker Expression Scoring

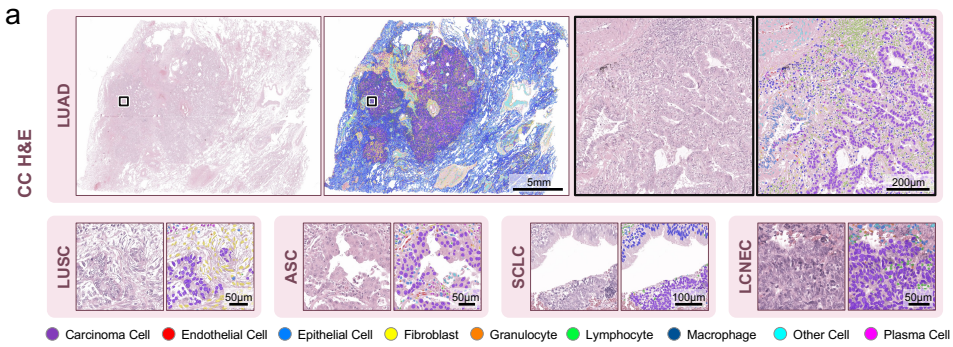
Building on the LUCAID IHC cell phenotyping module, we developed expression scoring modules to quantify biomarker expression intensity at individual cell level and subcellular compartments. Intensity was classified into four categories (negative, weak, moderate, and strong) for membranous MET and TROP-2 as well as cytoplasmic TROP-2 expression. PD-L1 was assessed using a binary classification scheme (positive vs. negative). Figure 2j–l presents examples of expert test annotations across the different expression scoring categories. For ADC (TROP-2 and MET) assessment, the test datasets included 18,959 (TROP-2 = 13,062 and MET = 5,897) and 2,678 annotations for membrane and cytoplasm, respectively (Figure 2a). For PD-L1, 5,650 annotations were collected for testing (Figure 2a). Module performance was highest for PD-L1, reaching an average F1 score of 0.93 (Figure 5d). For MET and TROP-2, F1 scores reached 0.87 for membranous and 0.82 for cytoplasmic expression across all intensity categories. Notably, classification performance for the clinically most relevant moderate and strong expression categories ranged between 0.86 and 0.93, supporting accurate identification of cases potentially eligible for targeted therapies. Overall, the consistent performance observed across distinct biomarkers and subcellular expression patterns highlights the strong generalizability of LUCAID expression scoring. This broad applicability is particularly relevant for emerging therapeutic targets, including the rapidly expanding class of ADC targets, where the same analytical framework can be applied across biomarkers with different staining patterns and scoring requirements.

## 2.2 Cellularity Estimates Correlate with Molecular Reference Data

To evaluate the clinical applicability of automated tumor cellularity assessment for molecular diagnostics, we analyzed a cohort of 115 *KRAS*-mutated lung cancer cases with available molecular profiling data. Tumor cellularity was quantified within pathologist-annotated tumor regions selected for molecular testing (Supplementary Figure 3a), and LUCAID-derived estimates and routine pathologist assessments were compared with *KRAS* variant allele frequency (VAF) as an orthogonal molecular reference for tumor-derived DNA content (see Methods: Comparison of Tumor Cellularity Assessment Variants). Routine pathologist assessment showed only weak correlation with the molecular reference ($r = 0.20$, $p = 0.033$), resulting in a mean absolute error (MAE) of 24.42 percentage points (ppt; Figure 5f). Using the conventional cell count-based approach, LUCAID improved correlation with *KRAS* VAF to $r = 0.39$ ($p < 0.001$) while reducing the MAE to 21.7 ppt (Figure 5g). Notably, the nuclear area-based tumor cellularity metric further strengthened agreement with the molecular reference, increasing the correlation to $r = 0.46$ ($p < 0.001$) and reducing the MAE to 17.1 ppt, corresponding to a 30.0% reduction relative to routine pathologist assessment and a 21.2% reduction relative to cell count-based quantification (Figure 5h; representative cell classification heatmaps within pathologist-annotated tumor regions are shown in Supplementary Figure 3b). Taken together, these findings demonstrate that automated cell-level analysis can improve the assessment of tumor cellularity for molecular profiling in a setting characterized by substantial interobserver variability among pathologists [53–55] and suggest that nuclear area-based quantification may provide a more accurate estimate of tumor cellularity than conventional cell count-based approaches.

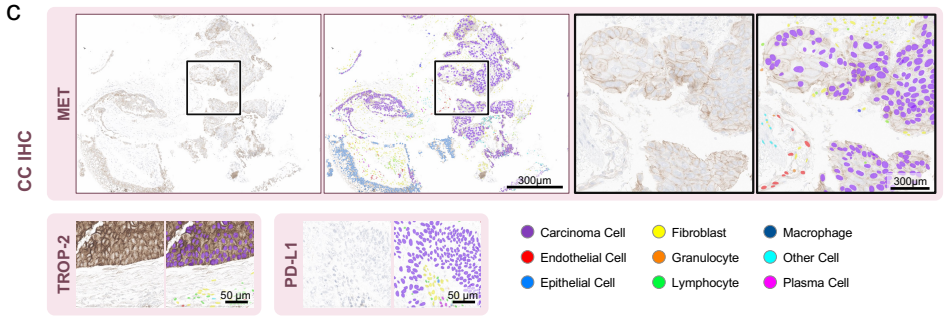
## 2.3 LUCAID Captures Patterns of Tumor Invasion and Identifies Prognostic TME Features

Beyond its core diagnostic modules, we investigated whether LUCAID-derived tissue- and cell-level measurements could capture clinically relevant histopathological patterns and extend routine assessment toward comprehensive characterization of the TME at tissue, cellular and spatial levels. By integrating tissue segmentation with cell phenotyping, LUCAID supported automated assessment of vascular invasion (Figure 6a), lymphatic invasion (Figure 6b) and spread through air spaces (STAS; Figure 6c), established prognostic features included in routine diagnostic reporting [56–58]. Quantification of viable tumor, necrosis, and stroma further enabled automated regression grading of resected specimens following neoadjuvant therapy according to the recommendations of the International Association for the Study of Lung Cancer (IASLC)(Figure 6d) [59]. LUCAID quantifications closely matched the joint pathologist assessment across all tissue compartments (pooled Pearson $r = 0.996$, $p < 0.001$; Spearman $\rho = 0.96$, $p < 0.001$), with mean absolute errors of 1.6, 1.9 and 2.9 percentage points for carcinoma, necrosis and stroma, respectively (Figure 6e,f). Thus, a semiquantitative visual estimate of pathological treatment response could be translated into a reproducible quantitative measurement. Applied to 1,001 patients of the discovery cohort with available follow-up, LUCAID quantified the distribution of tissue, cellular and spatial TME features across tumors (Figure 6g,h), including features that are too time-consuming for routine manual assessment or not practically quantifiable by visual inspection alone. These cohort-level distributions served as reference distributions for contextualizing individual LUCAID-derived measurements and subsequent report-level interpretation. UICC-adjusted Cox proportional-hazards analyses recapitulated established prognostic associations, including higher stromal tumor-infiltrating lymphocyte (TIL) levels with improved overall survival and higher tissue neutrophil-to-lymphocyte ratio (NLR) with adverse outcome (Figure 6i–k)[19, 20, 60]. Beyond these established markers, greater carcinoma cell–plasma cell adjacency was associated with improved survival, whereas greater endothelial cell–lymphocyte distance was associated with adverse outcome, revealing additional prognostic information in the spatial organization of the TME (Figure 6l,m). Finally,




b

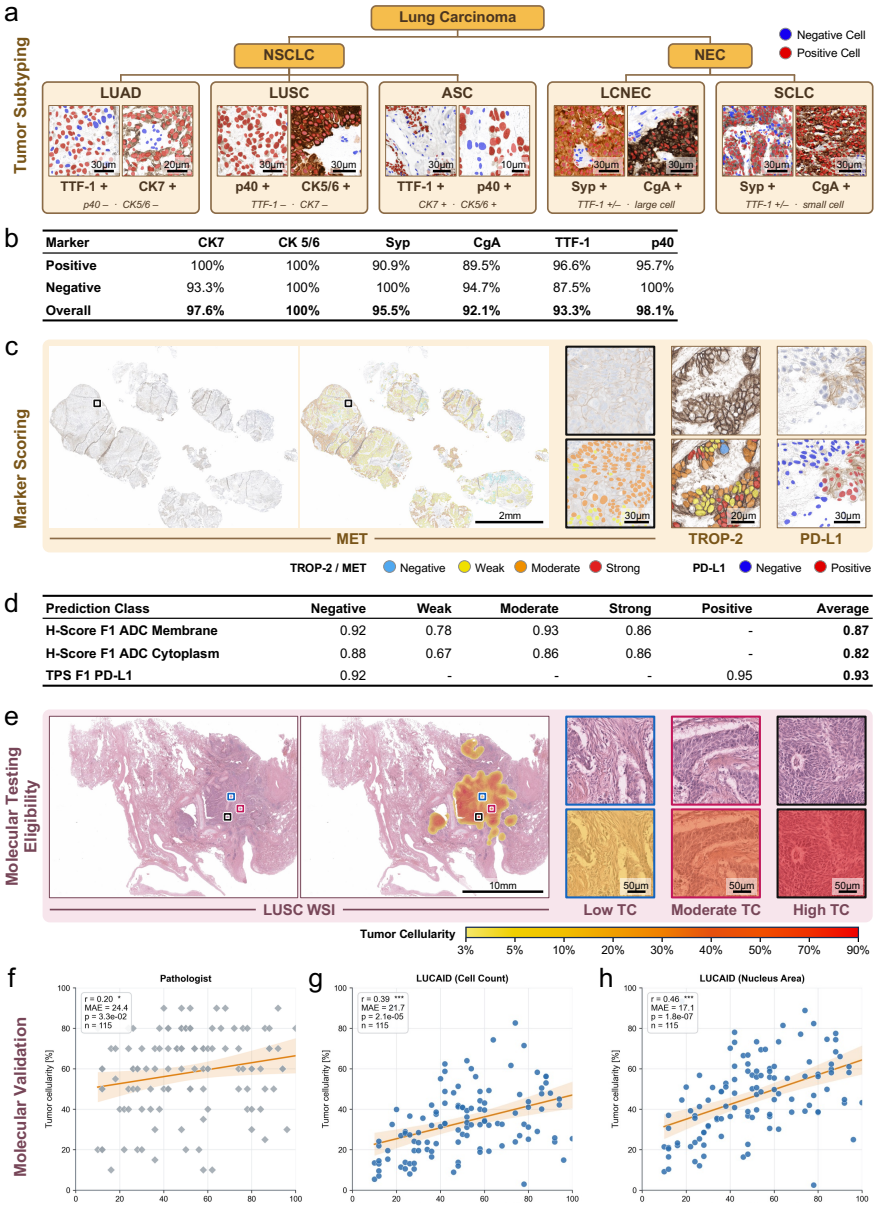
| Prediction Class | CaC | EC | Fib | Lym | PC | GC | M | EnC | Other Cell | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **F1 Overall** | 0.95 | 0.89 | 0.92 | 0.96 | 0.96 | 0.96 | 0.93 | 0.97 | 0.93 | **0.95** |
| **F1 LUAD** | 0.99 | 0.86 | 0.93 | 0.97 | 0.93 | 0.95 | 0.93 | 0.96 | 0.95 | **0.94** |
| **F1 LUSC** | 0.99 | 0.96 | 0.98 | 0.97 | 0.99 | 0.97 | 0.99 | 0.98 | 0.97 | **0.98** |
| **F1 ASC** | 1.00 | 1.00 | 0.96 | 0.94 | 0.87 | 0.96 | 0.98 | 0.97 | 0.98 | **0.96** |
| **F1 SCLC** | 0.99 | 0.93 | 0.94 | 0.87 | 0.97 | 0.93 | 0.68 | 0.94 | 0.87 | **0.90** |
| **F1 LCNEC** | 1.00 | 0.91 | 0.90 | 0.92 | 1.00 | 0.91 | 0.86 | 0.97 | 0.83 | **0.92** |




d

| Prediction Class | CaC | EC | Fib | Lym | PC | GC | M | EnC | Other Cell | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **F1 MET** | 0.97 | 0.68 | 0.96 | 0.97 | 0.94 | 0.93 | 0.88 | 0.78 | 0.60 | **0.86** |
| **F1 TROP-2** | 0.98 | 0.62 | 0.94 | 0.94 | 0.90 | 0.94 | 0.81 | 0.86 | 0.76 | **0.86** |
| **F1 PD-L1** | 0.95 | 0.71 | 0.96 | 0.95 | 0.89 | 0.95 | 0.91 | 0.93 | 0.89 | **0.90** |

Figure 4: **H&E and IHC cell phenotyping across lung cancer subtypes and biomarkers.**
**a**, H&E cell-phenotyping examples across five major lung carcinoma subtypes, with histopathological images shown on the left and corresponding model-derived cell-classification overlays on the right. For LUAD, a resection specimen is shown with a higher-magnification view; the lower row shows LUSC, ASC, SCLC and LCNEC from left to right. The model distinguishes nine cell classes: carcinoma cells, endothelial cells, epithelial cells, fibroblasts, granulocytes, lymphocytes, macrophages, plasma cells and other cells. **b**, F1 scores for each cell class overall and stratified by histological subtype. **c**, IHC cell-phenotyping examples for MET, TROP-2 and PD-L1 using the same nine cell classes, with histopathological images shown on the left and corresponding model-derived cell-classification overlays on the right. For MET, a biopsy is shown with a higher-magnification view; TROP-2 and PD-L1 are each shown as a single image–overlay pair. **d**, F1 scores for each cell class stratified by biomarker.
*Abbreviations:* ASC, adenosquamous carcinoma; CaC, carcinoma cell; CC, cell classification; EC, endothelial cell; EnC, epithelial cell; Fib, fibroblast; GC, granulocyte; H&E, hematoxylin and eosin; IHC, immunohistochemistry; LCNEC, large cell neuroendocrine carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; Lym, lymphocyte; M, macrophage; MET, hepatocyte growth factor receptor; PC, plasma cell; PD-L1, programmed death-ligand 1; SCLC, small cell lung cancer; TROP-2, trophoblast cell-surface antigen 2.




| Marker | CK7 | CK 5/6 | Syp | CgA | TTF-1 | p40 |
|---|---|---|---|---|---|---|
| **Positive** | 100% | 100% | 90.9% | 89.5% | 96.6% | 95.7% |
| **Negative** | 93.3% | 100% | 100% | 94.7% | 87.5% | 100% |
| **Overall** | **97.6%** | **100%** | **95.5%** | **92.1%** | **93.3%** | **98.1%** |

| Prediction Class | Negative | Weak | Moderate | Strong | Positive | Average |
|---|---|---|---|---|---|---|
| **H-Score F1 ADC Membrane** | 0.92 | 0.78 | 0.93 | 0.86 | - | **0.87** |
| **H-Score F1 ADC Cytoplasm** | 0.88 | 0.67 | 0.86 | 0.86 | - | **0.82** |
| **TPS F1 PD-L1** | 0.92 | - | - | - | 0.95 | **0.93** |

Figure 5: **Tumor subtyping, biomarker scoring and tumor cellularity for molecular profiling.**
**a**, IHC-based tumor subtyping across five major lung carcinoma entities using six diagnostic markers (CK7, CK5/6, synaptophysin, chromogranin A, TTF-1 and p40). For each entity, two key diagnostic stains are shown as IHC images with model-derived cell-classification overlays; marker-positive cells are shown in red and marker-negative cells in blue. **b**, Accuracy for positive, negative and overall classification for each diagnostic marker. **c**, Biomarker expression-scoring examples for MET, TROP-2 and PD-L1. For MET, a biopsy overview is shown with the model-derived overlay, followed by a higher-magnification image–overlay pair; TROP-2 and PD-L1 are each shown as paired histopathological images and model-derived overlays. MET and TROP-2 expression is classified as negative, weak, moderate or strong, whereas PD-L1 expression is classified as negative or positive. **d**, F1 scores for each expression class for membranous and cytoplasmic ADC-target scoring and PD-L1 tumor proportion scoring. **e**, Tumor cellularity assessment for molecular testing eligibility. A resection specimen is shown as an H&E whole-slide overview with the corresponding model-derived cellularity map, together with higher-magnification examples of regions with low, moderate and high tumor cellularity; histopathological images are shown above and corresponding overlays below. **f**–**h**, Orthogonal molecular validation of tumor cellularity estimates against *KRAS* variant allele frequency (n = 115). Correlation between *KRAS* variant allele frequency and tumor cellularity for **f**, routine pathologist estimates, **g**, AI-predicted tumor cell proportion and **h**, AI-predicted tumor nuclear area proportion. Pearson $r$ and Spearman $\rho$ (both two-tailed) and the MAE versus variant allele frequency are shown in each panel. Orange lines indicate linear fits; shaded bands indicate 95% confidence intervals.
*Abbreviations:* ADC, antibody–drug conjugate; ASC, adenosquamous carcinoma; CgA, chromogranin A; CK5/6, cytokeratin 5/6; CK7, cytokeratin 7; H&E, hematoxylin and eosin; IHC, immunohistochemistry; LCNEC, large cell neuroendocrine carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MAE, mean absolute error; MET, hepatocyte growth factor receptor; PD-L1, programmed death-ligand 1; SCLC, small cell lung cancer; SYP, synaptophysin; TC, tumor cellularity; TTF-1, thyroid transcription factor 1; TROP-2, trophoblast cell-surface antigen 2.

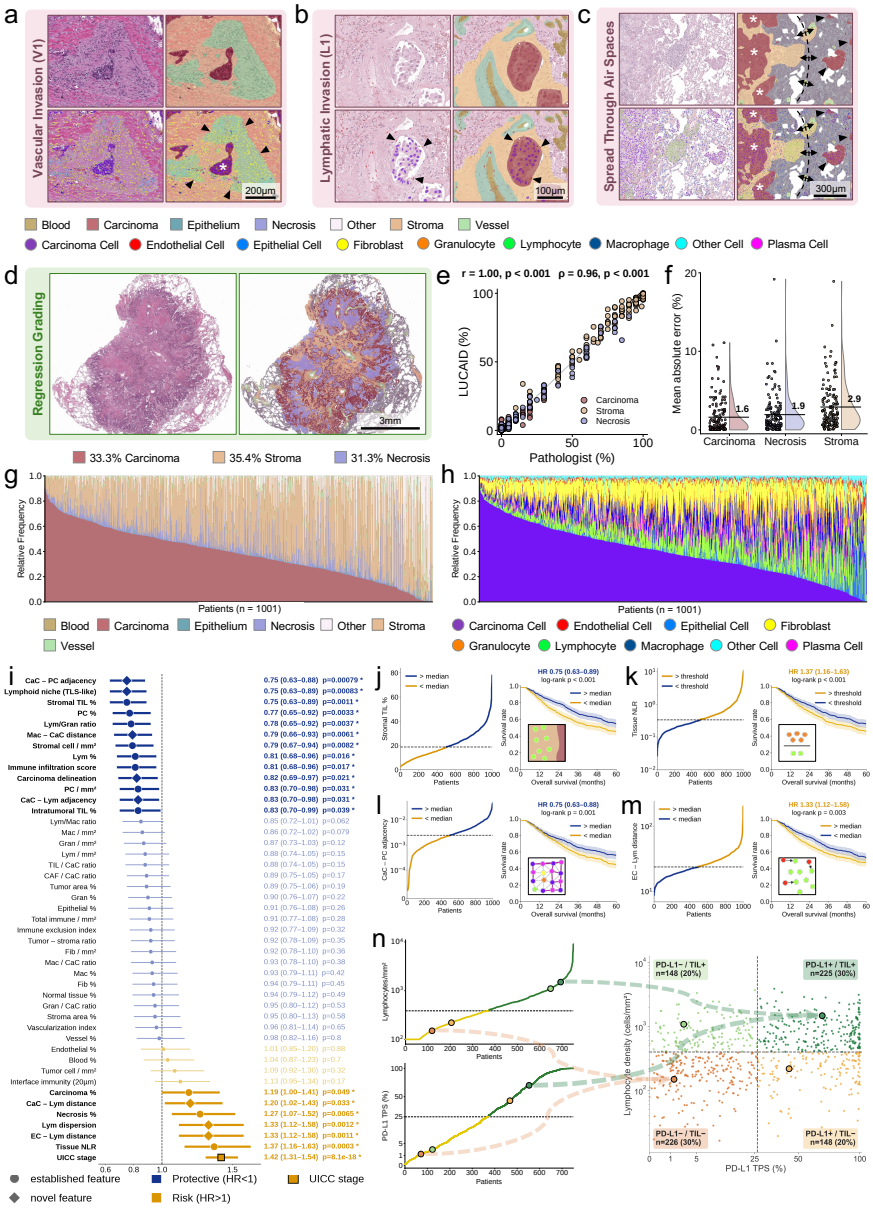



Figure 6: **Clinicopathological applications and biomarker discovery in lung cancer.**
**a**–**c**, Histopathological examples of vascular invasion (**a**), lymphatic invasion (**b**) and spread through air spaces (STAS; **c**), visualized using tissue-segmentation and cell-classification overlays. For each finding, H&E is shown at the top left, tissue segmentation at the top right, cell classification at the bottom left and the combined overlay at the bottom right. In vascular invasion, black arrows indicate the detected vessel and the white asterisk marks intravascular carcinoma. In lymphatic invasion, black arrows indicate endothelial cells surrounding the carcinoma cells. In STAS, white asterisks mark the main tumor, the dashed double-headed line indicates a defined distance from the tumor border within healthy lung parenchyma (gray), and black arrows indicate carcinoma within alveolar spaces beyond this zone. **d**, Pathological regression grading in a resection specimen, showing H&E (left) and the tissue-segmentation overlay (right) with quantification of viable carcinoma, stroma and necrosis. **e**, Correlation between the joint pathologist assessment and AI-predicted proportions of carcinoma, stroma and necrosis (n = 140). Pearson and Spearman correlation coefficients are shown; tests are two-sided. **f**, Mean absolute error (MAE) of AI-predicted carcinoma, stroma and necrosis proportions relative to the joint pathologist assessment. **g**, Discovery-cohort (n = 1,001) distribution of tissue-compartment composition across tumors, ordered by carcinoma proportion. **h**, Corresponding distribution of cell-type composition, ordered by carcinoma-cell proportion. **i**, UICC-adjusted Cox proportional-hazards analysis of tumor microenvironment (TME) features. Hazard ratios are shown as points with 95% confidence intervals; established features are shown as circles, novel spatial features as diamonds and UICC stage as a square. Protective and risk-associated features are shown in blue and orange, respectively; significant features ($p < 0.05$, log-rank test with Benjamini–Hochberg correction) are shown in bold. **j**–**m**, Cohort-level distributions and corresponding Kaplan–Meier overall-survival curves (log-rank test) stratified into high and low groups for **j**, stromal tumor-infiltrating lymphocyte (TIL) percentage; **k**, tissue neutrophil-to-lymphocyte ratio; **l**, carcinoma cell–plasma cell adjacency; and **m**, endothelial cell–lymphocyte distance. **n**, Cohort-level distributions of intratumoral lymphocyte density and PD-L1 tumor proportion score (TPS), with joint classification into four immune phenotypes defined by high or low TIL density and PD-L1 expression, which are associated with differential response to immune checkpoint inhibition.
*Abbreviations:* CaC, carcinoma cell; CAF, cancer-associated fibroblast; E, epithelial cell; Fib, fibroblast; Gran, granulocyte; H&E, hematoxylin and eosin; Lym, lymphocyte; Mac, macrophage; MAE, mean absolute error; NLR, tissue neutrophil-to-lymphocyte ratio; PC, plasma cell; PD-L1, programmed death-ligand 1; STAS, spread through air spaces; TIL, tumor-infiltrating lymphocyte; TME, tumor microenvironment; TPS, tumor proportion score; UICC, Union for International Cancer Control.

integrating intratumoral lymphocyte density with AI-predicted PD-L1 tumor proportion score (TPS) stratified tumors into four immune phenotypes defined by TIL-high/low and PD-L1-high/low status, mirroring a previously described TME classification associated with differential response to PD-1/PD-L1 blockade (Figure 6n) [61]. LUCAID provides a standardized and automated implementation of this immune phenotyping strategy by integrating H&E-derived intratumoral lymphocyte density with IHC-based PD-L1 assessment, making combined immune phenotyping scalable across larger cohorts.

Taken together, these analyses show that LUCAID spans established diagnostic pathology, quantitative treatment-response assessment and advanced TME characterization, including invasion patterns, pathological regression grading, established prognostic markers, novel spatial features associated with outcome, and combined immune phenotypes associated with differential treatment response. The resulting cohort-wide distributions provide a reference for case-level interpretation and LUCAID report generation (Figure 7) and serve as a benchmark for future studies of the lung TME.

## 2.4 LUCAID Generates Structured Pathology Reports Grounded in Validated Modules

The outputs of LUCAID modules provide a broad set of quantitative biological measurements that require clinical interpretation and accessible presentation for routine use. We therefore evaluated whether LUCAID could integrate these validated module outputs into structured pathology reports and whether the generated content remained faithful to the underlying measurements.

For a given case, the outputs of all applicable modules are retrieved and assembled into a report organized into predefined diagnostic sections. Each section pairs the quantitative measurements of one module with a concise interpretation of their diagnostic or therapeutic relevance. A concluding case-level assessment integrates findings across modules and summarizes their potential clinical implications and treatment considerations. Quantitative values are transferred directly from the validated analysis modules, separating the generation of measurements from their subsequent LUCAID-based clinical contextualization and interpretation.

To ground these interpretations in published evidence, we implemented a PubMed retrieval and citation mechanism. Before generating an interpretation, the agent queries PubMed [+] through the NCBI E-utilities interface and and retrieves potentially relevant publications. The agent may then use these records to support selected interpretive statements. Crucially, citation entries are taken directly from PubMed, and LUCAID only selects which of the retrieved records to cite. References to non-existent publications therefore cannot appear in the report by construction. A representative report is shown in Figure 7a, with the corresponding reference list provided in Appendix B. Beyond the static report, the agent supports case-specific follow-up queries, calling relevant analysis modules as required and integrating their outputs into the response (example conversation shown in Figure 7b).

To assess report fidelity and clinical validity, we adapted and combined established evaluation frameworks [62–65] and decomposed generated reports into atomic statements. Board-certified pathologists independently evaluated each statement according to its function. Statements that directly reference a module output are graded for groundedness, defined as preservation of the direction and magnitude of the underlying measurement (grounded, partially grounded, ungrounded). Statements providing non-trivial clinical interpretations of one or more module outputs were graded for interpretation correctness. Statements carrying a citation are additionally graded for citation correctness, assessing whether the cited publication supported the associated claim (yes, partially, no). Incorrect statements were further assessed for potential harm (none, mild-to-moderate or severe) and for the likelihood of influencing a clinical decision under routine review (low, medium or high). At the report level, we

[+] https://pubmed.ncbi.nlm.nih.gov/

evaluated clinically relevant omissions and contradictions between section-level findings and the overall assessment. Detailed evaluation criteria and instructions are provided in Appendix B.

Ten reports containing outputs from all LUCAID modules were evaluated, yielding 1,365 atomic statements (median, 134.5 statements per report; range, 119–154). Of these, 134 (9.8%) directly referenced a module output and 1,231 (90.2%) contained a clinical interpretation based on one or more outputs (Figure 7c). All 134 referencing statements were graded as grounded, indicating that the agent reliably anchors its statements in the module outputs rather than introducing quantities of its own. Of the 1,231 interpretive statements, 1,194 (97.0%) were judged correct and 37 (3.0%) incorrect, distributed over nine of the ten reports. A total of 159 citations were placed, of which 131 (82.4%) fully or partially supported the associated statement, whereas 28 (17.6%) did not. At the report level, no clinically relevant omissions were identified across 70 section-level assessments; however, three contradictions between the overall assessment and individual report sections occurred in two reports (Figure 7c). Harm was rated for 111 errors. Of these, 96 (86.5%) were considered to have no potential for harm and 15 (13.5%) mild-to-moderate potential, with none rated as severe.

## 2.5 LUCAID Shows High Concordance across Clinically Actionable Tasks

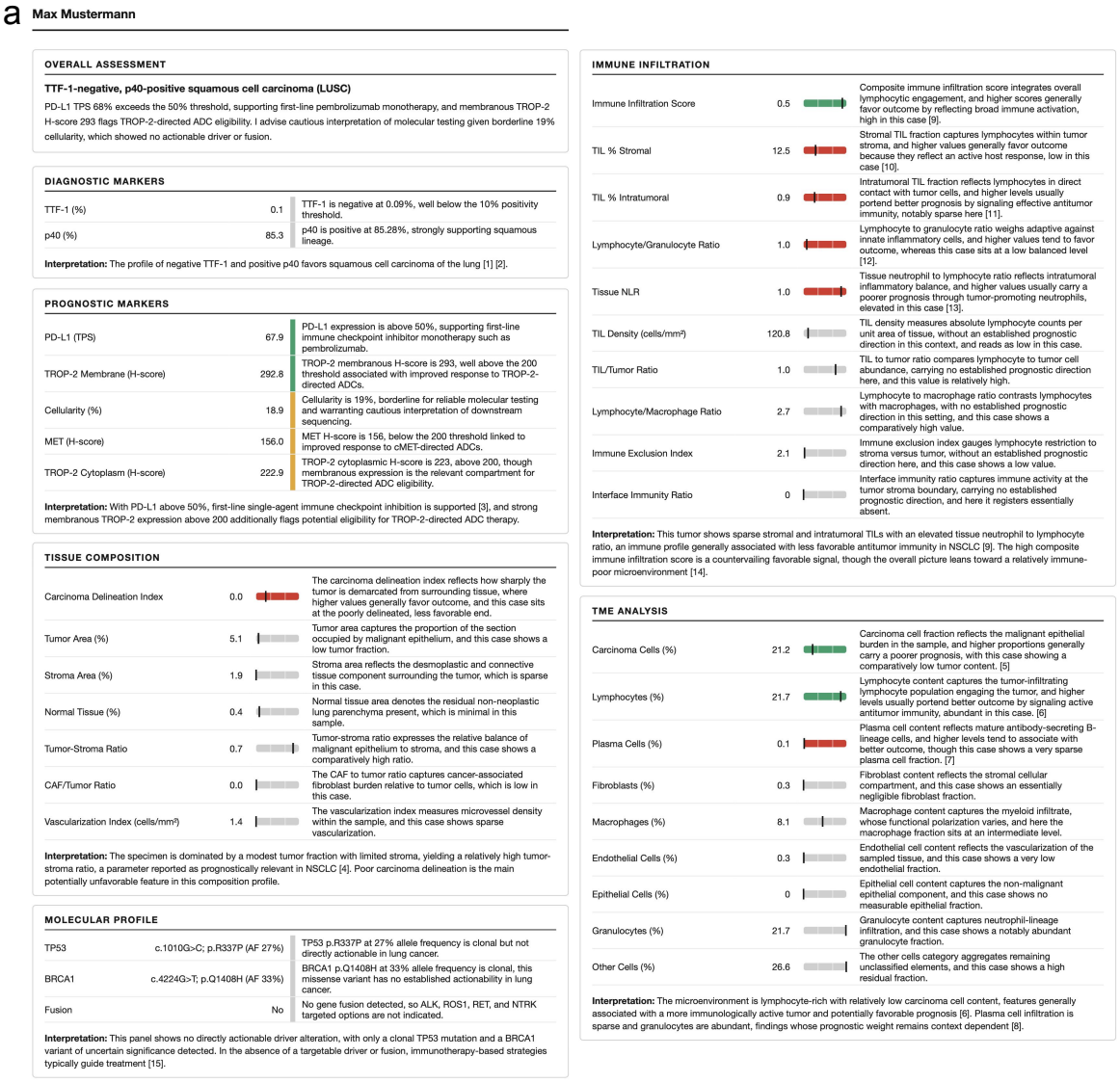
After validating the individual LUCAID modules against expert annotations, we assessed their performance on five key tasks in lung cancer diagnostics: tumor cellularity assessment for molecular testing, PD-L1 tumor proportion scoring (TPS), MET H-score assessment, and membranous and cytoplasmic TROP-2 H-score assessment. Seventy consecutive patients with lung cancer were prospectively enrolled through the nNGM program (cohort characteristics are summarized in Supplementary Table 1).

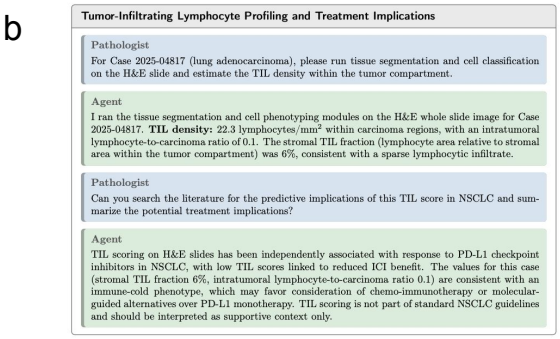
For each case, the five clinically actionable diagnostic tasks were assessed by the corresponding LUCAID modules and benchmarked, task by task, against five experienced pathologists across 338 diagnostic assessments (clinical concordance; cf. Section 4.10). An expert-panel adjudicated reference standard, established after a three-month washout period, served as the reference for all comparisons (cf. Section 4.9).

Continuous scores showed strong correlations with the reference for LUCAID and all five pathologists across the five tasks, with all correlations reaching statistical significance (all two-sided $p < 0.001$; Figure 8a). Across the six raters, LUCAID ranked third overall by correlation strength, indicating broadly comparable agreement at the continuous-score level despite variability between individual raters (Supplementary Figure 4a). A representative lung adenocarcinoma case illustrates the corresponding LUCAID assessments across the diagnostic workflow (Figure 8b).

Differences between evaluators became more apparent when considering quantitative error. LUCAID showed the lowest mean absolute error (MAE) for tumor cellularity assessment (3.46 percentage points; $\rho = 0.975$), compared with 8.9–21.23 percentage points for the pathologists. The corresponding MAEs were 6.70 percentage points for PD-L1 ($\rho = 0.959$), 19.11 for MET ($\rho = 0.926$), 20.82 for membranous TROP-2 ($\rho = 0.776$) and 24.20 for cytoplasmic TROP-2 ($\rho = 0.846$), with LUCAID ranking among the lowest-error raters across tasks (Figure 8a,c). For tumor cellularity, evaluation against *KRAS* variant allele frequency (VAF) as an orthogonal molecular reference further supported the nucleus area-based approach (MAE = 10.91 percentage points, $r = 0.776$), compared with the tumor cell count-based approach (MAE = 18.25 percentage points, $r = 0.760$) and pathologist estimates (MAE = 14.89–23.68 percentage points; $r = 0.18$–$0.70$; Supplementary Figure 4b).

At clinically relevant decision thresholds, LUCAID achieved an overall clinical action concordance of 93.0% with the expert-panel adjudicated reference standard, compared with 68.3–81.1% for the individual pathologists (Figure 8d). Concordance was 98.6% for molecular testing eligibility based on tumor cellularity ([0–10%) versus [10–100%]), compared with 89.9–92.9% for the pathologists; 89.9% for PD-L1




c

| | Referencing | | Interpretive | | Citations | | Report level | |
|---|---|---|---|---|---|---|---|---|
| Report | $N$ | % | $N$ | % | $N$ | % | Om. | Con. |
| R1 | 16 | 100 | 136 | 97.8 | 15 | 86.7 | 0 | 0 |
| R2 | 22 | 100 | 129 | 100 | 19 | 84.2 | 0 | 0 |
| R3 | 18 | 100 | 136 | 97.1 | 13 | 92.3 | 0 | 0 |
| R4 | 15 | 100 | 124 | 96.8 | 16 | 75.0 | 0 | 0 |
| R5 | 6 | 100 | 119 | 94.1 | 15 | 80.0 | 0 | 1 |
| R6 | 13 | 100 | 117 | 98.3 | 16 | 87.5 | 0 | 0 |
| R7 | 7 | 100 | 120 | 95.0 | 15 | 80.0 | 0 | 0 |
| R8 | 13 | 100 | 116 | 94.8 | 17 | 76.5 | 0 | 0 |
| R9 | 10 | 100 | 109 | 99.1 | 18 | 72.2 | 0 | 0 |
| R10 | 14 | 100 | 125 | 96.8 | 15 | 93.3 | 0 | 2 |
| **All** | **134** | **100** | **1231** | **97.0** | **159** | **82.4** | **0** | **3** |

Figure 7: **Structured report generation, agentic interaction and report evaluation.**
**a**, Representative structured diagnostic report generated by LUCAID. Module outputs are integrated into predefined sections and interpreted for clinical relevance. For tissue-composition and TME readouts without established clinical cut-offs, values are contextualized by their percentile position within reference distributions derived from the discovery cohort. LUCAID integrates the section-level findings into an overall case assessment. **b**, Representative pathologist–agent dialogue in natural language. The pathologist requests TIL density estimation and potential treatment implications; LUCAID provides an evidence-grounded answer using quantitative AI module results and retrieved literature. **c**, Evaluation of ten generated reports at claim and report level. Reports were decomposed into atomic statements. Referencing denotes statements that directly cite quantitative module results; the percentage indicates the proportion judged grounded in direction and magnitude of the result. Interpretive denotes statements drawing a clinical inference from one or more module results; the percentage indicates the proportion judged correct. Citations reports the number of literature citations and the proportion judged to (partially) support the associated statement. At the report level, omissions (Om.) indicate clinically relevant findings evident from the reported results but not reflected in the summary sections. Contradictions (Con.) indicate inconsistencies between the overall assessment and the individual report sections.
*Abbreviations:* ADC, antibody–drug conjugate; Con., contradiction; IHC, immunohistochemical; MET, hepatocyte growth factor receptor; NSCLC, non-small cell lung cancer; Om., omission; PD-L1, programmed death-ligand 1; QC, quality control; TIL, tumor-infiltrating lymphocyte; TME, tumor microenvironment; TPS, tumor proportion score; TROP-2, trophoblast cell-surface antigen 2; TTF-1, thyroid transcription factor 1.

treatment stratification (TPS [0–1%), [1–50%) and [50–100%]), compared with 70.1–76.8%; and 92.5% for MET classification (H-score [0–100), [100–200) and [200–300]), compared with 56.7–75.8%. The greatest interobserver variability was observed for TROP-2, with concordance of 95.2% for membranous expression versus 48.4–93.5% among pathologists and 88.7% for cytoplasmic expression versus 50.0–74.2% (Figure 8d). Case-level analysis further demonstrated substantial variation in agreement across the prospective cohort. Overall, 122 assessments (36%) differed between evaluators, with the extent of disagreement varying markedly between individual cases (Figure 8e). Discrepant cases revealed clinically relevant threshold effects. In low-cellularity tumors, pathologists more often estimated tumor content above the 10% threshold, whereas LUCAID remained below it (Supplementary Figure 5a,b), consistent with reports of visual overestimation of tumor fractions [53–55]; such overestimation could increase the risk of false-negative molecular testing. For PD-L1, LUCAID identified small populations of positive tumor cells crossing the 1% threshold that were missed visually, potentially affecting eligibility for immune checkpoint inhibitor therapy (Supplementary Figure 5c). Similar threshold-related discrepancies were observed for MET and TROP-2 (Supplementary Figure 5d–f). In summary, prospective clinical validation in a real-world precision oncology setting showed consistently high LUCAID concordance with the expert-panel adjudicated reference standard across all five clinically actionable diagnostic tasks. These findings highlight the potential of standardized AI-assisted assessment to improve consistency in borderline cases close to clinically relevant decision thresholds and to reduce recurrent observer-related variation in routine diagnostic assessment.

# 3 Discussion

Precision oncology in lung cancer increasingly depends on the integrative evaluation of spatial histological features and complex multimodal biomarkers, yet this often exceeds the time and cognitive limits of routine assessment, leaving much of the available information out of clinical decisions. AI can resolve this problem at scale, but the field has lacked models that reliably achieve expert-level performance and systems that span and integrate the complete diagnostic workflow rather than individual steps. To close this gap, we developed a suite of individual AI modules that cover each step of the diagnostic process and achieve pathologist-level accuracy across all major histological lung cancer subtypes. We then introduced an agent that orchestrates these models to form LUCAID, an agentic system that lets pathologists run complex multi-step analyses through natural language, interpret the findings in biological context, and generate structured pathology reports. In prospective, multicenter validation on a real-world lung cancer cohort, LUCAID's end-to-end execution outperformed five experienced thoracic pathologists across all clinically actionable decision tasks, reaching 93.0% agreement with the adjudicated expert reference standard versus 68.3–81.1% for individual pathologists.

At the core of LUCAID lies a collection of modules that were all validated against large-scale expert annotations and together cover the complete diagnostic workflow, from tissue quality and tumor detection through subtyping, TME characterization, and tumor cellularity assessment to subcellular biomarker scoring. Related prior approaches such as GrandQC [48] for quality control, the nnU-Net-based segmentation models of Kludt et al. [66] and Spronck et al. [67], or HistoPLUS [68] for H&E cell phenotyping address only isolated steps of the pathology workflow and none has been prospectively validated against pathologists on the threshold-based decisions that determine treatment. The combination of task coverage and performance that the LUCAID modules provide is unprecedented in lung cancer and we show that this unlocks clinically relevant applications that had previously remained out of reach.

Firstly, it enables us to perform all clinically actionable decision tasks that guide systemic therapy selection in lung cancer, such as assessing molecular testing eligibility and PD-L1 tumor proportion

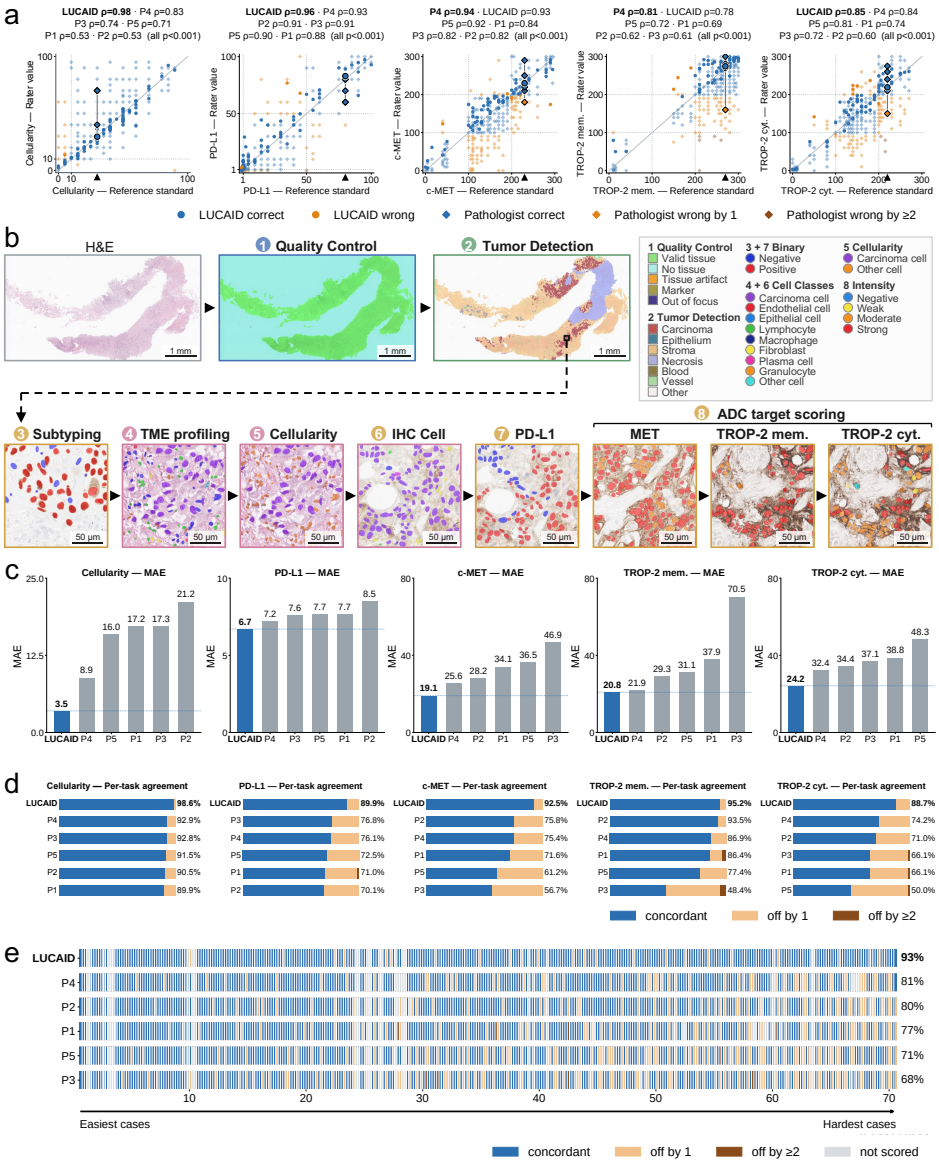



Figure 8: **Prospective clinical validation against an expert-panel adjudicated reference standard.** **a**, Rater-versus-reference comparisons for LUCAID and five pathologists (P1–P5) across tumor cellularity, PD-L1 tumor proportion score (TPS), MET H-score, TROP-2 membranous H-score and TROP-2 cytoplasmic H-score (numbers of evaluable cases per task in Section 4.1.3). Dashed lines indicate clinical decision thresholds. Spearman correlation coefficients are shown for each rater; all tests are two-sided. **b**, Representative LUAD case showing LUCAID outputs across the diagnostic workflow, including quality control, tumor detection, tumor subtyping, TME profiling, tumor cellularity assessment, IHC cell phenotyping, PD-L1 scoring and MET/TROP-2 scoring. Histopathological images are shown with the corresponding model-derived overlays. **c**, Mean absolute error (MAE) relative to the expert-panel adjudicated reference standard for LUCAID and each pathologist across the five scoring tasks, ranked from lowest to highest error within each task. **d**, Clinical action concordance for LUCAID and each pathologist, defined as the proportion of cases assigned to the same clinically relevant category as the reference standard. Thresholds were tumor cellularity <10% versus ≥10%; PD-L1 TPS <1%, 1–49% and ≥50%; and MET and TROP-2 H-scores <100, 100–199 and ≥200. Bars indicate concordant classifications and deviations from the reference by one or at least two categories. **e**, Case-level clinical action concordance across 70 prospectively evaluated cases, ordered from highest to lowest overall concordance. Each case is represented by five tiles corresponding to the five scoring tasks; colors indicate concordance, deviations from the reference by one or at least two categories, and tasks not scored.

*Abbreviations:* H&E, hematoxylin and eosin; IHC, immunohistochemical; MAE, mean absolute error; MET, hepatocyte growth factor receptor; PD-L1, programmed death-ligand 1; TME, tumor microenvironment; TPS, tumor proportion score; TROP-2, trophoblast cell-surface antigen 2.

scoring, in a standardized, quantitative and reproducible manner. This stands in stark contrast to conventional semi-quantitative visual assessment by a pathologist, which has been shown to suffer from substantial intra- and interobserver variability [27, 29, 69]. This is further underlined by the results on our prospective cohort, where the five pathologists disagreed on the clinical action category in 122 of 338 assessments (Figure 8). More broadly, this suggests, that perhaps the principal benefit of AI for the field may not be automation, but rather the standardization and reproducibility of complex diagnostic decisions across institutions and observers.

Secondly, the modules generalize across histological types, focus on biological primitives such as cell phenotypes and tissue compartments, and readily provide measurements at cohort scale. This enables both validation and discovery of biomarkers that LUCAID has never been explicitly trained for. Importantly, this includes routine applications such as pathological regression grading after neoadjuvant therapy that can directly be performed based on the tissue compartment scores and where we show that LUCAID reproduces the joint pathologist assessment of viable tumor, necrosis and stroma to within 1.6–2.9 percentage points (Figure 6d-f). Furthermore, it also opens up new assessments that are not yet established in clinical practice because they are too laborious or too fine-grained for visual evaluation. For instance, combining the intratumoral lymphocyte density from the cell phenotyping module with the PD-L1 score from the expression scoring module allows for the stratification of patients into four immune phenotypes associated with differential response to checkpoint inhibition (Figure 6n). Applications like this have so far remained research constructs as they require the reliable quantification of multiple markers at scale. We also show that for all these biomarkers, we can compute reference distributions across 1,001 lung cancer patients, which empirically contextualize the measurements for an individual case within the cohort. This can be used for discovery to identify candidate prognostic features (Figure 6i), such as carcinoma cell–plasma cell adjacency and endothelial cell–lymphocyte distance, but also as a help for practitioners to interpret readouts without established clinical cut-offs. We illustrate the latter, by visualizing the reference distributions in the generated reports (Figure 7). All in all, the broad variety of applications that LUCAID supports by quantifying fundamental biological building blocks of tissue, suggests that one way forward for computational pathology could be to focus on developing models that comprehensively and reliably estimate these quantities across indications and institutions. This would stand in contrast to the task-specific approaches that have so far been omnipresent in the field [70–72], and would shift the emphasis for practitioners from endpoint-specific model development towards deriving biological insight from reliable, generalizable readouts.

Thirdly, complete workflow coverage is what allows the agent to assemble a structured report that integrates all relevant parameters of a case and covers the complete diagnostic workflow (Figure 7). The reports LUCAID generates are grounded in a stricter sense than previous work has been able to achieve. Systems that generate reports from whole-slide images generally reason directly over the image with a language or vision–language model [42, 73, 74], which allows flexible conversation but is ill-suited to the threshold-based quantitative decisions at the core of precision oncology. The tumor proportion score or staining intensity per subcellular compartment require counting and density estimation across the entire slide, while vision–language models operate on limited fields of view and do not perform reliably on fine-grained spatial tasks [44–47]. And while the need for grounding and reliability is widely recognized, existing approaches ground statements in visual or literature evidence rather than in standardized biological metrics with known performance characteristics. For example, QCAgent verifies its statements by re-retrieving the slide regions that support them, and PathPocket traces interpretations to a corpus of published evidence [75, 76]. Both cases attempt to make the evidence retrievable, but neither reliably quantifies the underlying biology. LUCAID instead builds the report directly on validated measurements. The agent contextualizes the module outputs against established clinical thresholds and provides the diagnostic and therapeutic interpretation, but the values in the report are filled in directly from the modules that produced them. Failure modes that generally make

generated reports hard to trust, such as the hallucination of plausible biomarker values from a diagnostic label, are excluded by construction.

This is underlined by our preliminary demonstration, in which board-certified pathologists decomposed ten generated reports into 1,365 atomic statements. Of these, 134 (9.8%) referenced a module output directly and all were faithful to the underlying value. Furthermore, of the 1,231 (90.2%) that drew a clinical interpretation, 1,194 (97.0%) were judged correct. Among errors rated for harm, 15 (13.5%) carried mild-to-moderate potential and none was rated severe. This gives a first indication that restricting text generation to the bounds set by validated module measurements may be a path toward reliable language model use in clinical reporting. Additionally, as LUCAID's language-based interpretation is separate from the measurements printed beside it, a practitioner can recognize faulty reasoning from the report itself. Citation placement was assessed in the same way, with 131 of 159 citations (82.4%) supporting or partially supporting the statement they accompanied. Since entries are taken from PubMed, fabricated references cannot occur by construction, though refining which of the retrieved records the agent selects and adding additional literature data-bases are natural next steps.

Our approach also has limitations. It focuses on lung cancer, and extending the modules to further indications is a logical next step. While the system was validated on a multicentric cohort and prospectively against pathologists, its influence on real-world clinical decision-making, and how pathologists would interact with it in routine practice, remain to be established. Similarly, the evaluation of the generated reports rests on ten cases and should be read as a proof of principle rather than a full-scale clinical validation. Finally, the orchestrating agent is built on a general-purpose model and domain-specific alternatives, fine-tuning, and refinement with human feedback are straightforward routes to further gains.

Looking beyond these open questions, our work points toward a broader vision for how AI might be integrated into pathology, and takes a concrete step in that direction. For AI to earn trust in clinical settings, it must perform sequential analyses across the entire whole-slide image that are consistent between patients and interpretable by a pathologist. Routine adoption additionally depends on a natural language interface, which allows such systems to integrate into pathology workflows and remain accessible to non-technical users. LUCAID was designed to meet both requirements, pairing precise and reproducible cell-level quantitative analysis across the whole slide with a conversational interface that a pathologist can direct without technical expertise. Taking a glimpse into the future, we envision a marketplace of validated pathology tools, each covering a clinically important part of the diagnostic process and independently approved for clinical use, from which a pathologist assembles and directs those required for a given case through natural language. The present work realizes this principle at the example of lung cancer diagnostics, showing how reliable quantitative AI can be brought to the full diagnostic workflow.

# 4 Methods

## 4.1 Clinical Cohorts

Cohorts were assembled to maximize morphological and technical heterogeneity and thereby support model generalization across tissue types, staining modalities, and imaging domains.

#### 4.1.1 Test Cohorts for Quality Control, Tissue Segmentation, Cell Phenotyping and Biomarker Scoring

The H&E tissue segmentation and cell phenotyping test cohort consisted of 158 primary lung tumor WSIs ($n = 117$ from Charité – University Medical Center Berlin and $n = 41$ from external referring centers) and 47 metastic lung tumors ($n = 39$ from Charité – University Medical Center Berlin and $n = 8$ from external referring centers) scanned across two scanner platforms (Aperio GT 450 DX and Aperio AT2 (Leica Biosystems)). The H&E QC cohort comprised 132 WSI across ($n = 83$ from Charité – University Medical Center Berlin and $n = 49$ from external referring centers) across two scanner platforms (Aperio GT 450 DX and Aperio AT2 (Leica Biosystems)), while the IHC QC cohort comprised 106 WSIs ($n = 52$ from Charité – University Medical Center Berlin and $n = 54$ from external referring centers) representing a broad spectrum of tumor entities, including lung carcinoma ($n = 72$). WSIs were scanned across eleven scanner platforms (Aperio GT 450 DX, Aperio AT2 and Aperio ScanScope (Leica Biosystems); VENTANA DP 200, VENTANA DP 600 and VENTANA iScan (Roche Diagnostics); Pannoramic 1000, Pannoramic 250 Flash III, Pannoramic SCAN II and Pannoramic MIDI II (3DHISTECH); and Vectra Polaris (Akoya Biosciences)).

For IHC cell phenotyping and biomarker scoring, primary lung carcinoma cases were scanned on two platforms (Pannoramic 1000 (3DHISTECH); VENTANA DP 600 (Roche Diagnostics)). Cases were distributed across tasks as follows: TROP-2 cell phenotyping, $n = 32$ ($n = 22$ from Charité – University Medical Center Berlin and $n = 10$ from University Hospital Cologne); MET cell phenotyping, $n = 30$ ($n = 20$ from Charité – University Medical Center Berlin and $n = 10$ from University Hospital Cologne); PD-L1 cell phenotyping, $n = 35$ ($n = 30$ from Charité – University Medical Center Berlin and $n = 5$ from University Hospital Cologne); TROP-2 cytoplasmic expression scoring, $n = 29$ ($n = 19$ from Charité – University Medical Center Berlin and $n = 10$ from University Hospital Cologne); membranous expression scoring, $n = 59$ (TROP-2, $n = 29$: $n = 19$ from Charité – University Medical Center Berlin and $n = 10$ from University Hospital Cologne; MET, $n = 30$: $n = 20$ from Charité – University Medical Center Berlin and $n = 10$ from University Hospital Cologne); and PD-L1 expression scoring, $n = 35$ ($n = 30$ from Charité – University Medical Center Berlin and $n = 5$ from University Hospital Cologne).

#### 4.1.2 Discovery Cohort

The discovery cohort was a bicentric NSCLC cohort of 1,001 surgically resected tumors collected between 2006 and 2019 at Charité – University Medical Center Berlin and University Hospital Cologne, comprising 581 lung adenocarcinomas (LUAD), 402 lung squamous cell carcinomas (LUSC) and 18 adenosquamous carcinomas (ASC). Clinicopathological characteristics are summarized in Supplementary Table 2.

#### 4.1.3 Prospective Clinical Validation Cohort

For prospective clinical validation, 70 consecutive lung cancer cases enrolled in the nNGM program were prospectively analyzed between November 2024 and March 2025. Cases originated from the Institute of Pathology, Charité – University Medical Center Berlin ($n = 16$), the Department of Neuropathology, Charité – University Medical Center Berlin ($n = 2$), and five referring pathology sites in Berlin and Brandenburg, Germany: MVZ Pathologie Berlin-Buch ($n = 25$), DRK Kliniken Berlin Westend ($n = 13$), Ernst von Bergmann Klinikum, Potsdam ($n = 7$), Carl-Thiem-Klinikum, Cottbus ($n = 5$), and Pathologie Spandau, Berlin ($n = 2$). Clinicopathological characteristics are summarized in Supplementary Table 1. Comprehensive molecular profiling according to the nNGM diagnostic workflow was performed for all cases (see Molecular Analysis section for further details).

PD-L1 staining was available for analysis in 67 cases, MET in 66 cases, and TROP-2 in 62 cases.

## 4.2 Immunohistochemical Tumor Classification

For immunohistochemical lung cancer classification, a marker panel comprising Thyroid Transcription Factor-1 (TTF-1), p40, cytokeratin 7 (CK7), cytokeratin 5/6 (CK5/6), synaptophysin (SYP), and chromogranin A (CgA) stained at the Ludwig Maximilian University Munich was assembled. A total of 274 marker-specific WSIs were available for analysis, encompassing both positive and negative staining patterns for each marker. Marker-specific datasets included 45 TTF-1 stained WSIs (29 positive, 16 negative), 54 p40 stained WSIs (23 positive, 31 negative), 41 CK7 stained WSIs (26 positive, 15 negative), 52 CK5/6 stained WSIs (24 positive, 28 negative), 44 SYP stained WSIs (22 positive, 22 negative), and 38 CgA stained WSIs (19 positive, 19 negative). Cytoplasmic and membranous staining patterns were evaluated for CK7, CK5/6, synaptophysin, and chromogranin A, whereas nuclear expression was assessed for TTF-1 and p40. Cases with $\geq 10\%$ positive tumor cells were classified as positive for the respective marker. Marker combinations were integrated according to established diagnostic criteria to assign histological subtypes, including LUAD, LUSC, Adenosquamous carcinoma (ASC), Large-cell neuroendocrine carcinoma (LCNEC), and Small-cell lung carcinoma (SCLC). LUAD was immunohistochemically defined by CK7 positivity, absence of p40, absent or limited CK5/6 expression, with or without TTF-1 expression. LUSC was characterized by p40 and CK5/6 positivity together with absent or limited CK7 staining. ASC was defined by the presence of both LUAD and LUSC components, with each component representing $\geq 10\%$ of the tumor. LCNEC was defined by large-cell morphology and positivity for at least one neuroendocrine marker (SYP or CgA), while lacking p40 and showing absent or limited CK7 staining. SCLC was defined by small-cell morphology and positivity for at least one neuroendocrine marker, together with absence of p40 and CK5/6 and absent or limited CK7 staining [52, 77–79].

## 4.3 Comparison of Tumor Cellularity Assessment Variants

In routine molecular pathology workflows, tumor-rich tissue regions are selected and annotated on tissue slides prior to sequencing, and tumor cellularity is estimated by pathologists as a surrogate for the fraction of tumor-derived DNA within the analyzed sample. Conventionally, tumor cellularity is determined as the proportion of tumor cells among all nucleated cells within the selected region. However, this approach assumes that tumor and non-neoplastic cells contribute equally to the overall DNA content independent of differences in nuclear size. We therefore evaluated an alternative nuclear area-based approach, based on the hypothesis that the relative nuclear area occupied by tumor cells may better approximate tumor-derived DNA content than cell counts alone. To compare the biological relevance of both approaches, *KRAS* variant allele frequency (VAF) was used as an orthogonal molecular reference for tumor DNA content. As oncogenic *KRAS* mutations are frequently clonal and typically heterozygous, the observed VAF is expected to approximate half of the tumor cell fraction under idealized conditions (for example, a *KRAS* VAF of 20% corresponding to approximately 40% tumor cellularity), with the limitation that this relationship is affected by copy-number alterations, allelic imbalance, subclonality, tumor heterogeneity, and technical factors. For this analysis, H&E-stained WSIs from 115 lung cancer patients enrolled in the nNGM program were collected at Charité – University Medical Center Berlin, Germany, between January 2020 and December 2024. Routinely assessed diagnostic WSIs were used, in which tumor-rich regions had been annotated by board-certified pathologists and tumor cellularity had been estimated for each selected region prior to molecular testing. Tumor cellularity was subsequently quantified by the AI pipeline within the same annotated tissue regions using two complementary approaches: (1) the proportion of carcinoma cells among all detected cells

and (2) the proportion of carcinoma nuclear area relative to the total nuclear area of all detected cells. AI-derived estimates and routine pathologist assessments were compared with *KRAS* VAF using Spearman correlation coefficients and mean absolute error (MAE; see Statistical Analysis section for further details).

## 4.4 Immunohistochemical Analysis

Immunohistochemical staining was performed using a BenchMark XT automated immunostainer (Ventana Medical Systems). Antigen retrieval was conducted using either CC1 mild buffer (Ventana Medical Systems) or ER2 buffer (Leica Biosystems) at 100 °C for 16-64 minutes, according to antibody-specific requirements. Slides were incubated with primary antibodies at room temperature for 60 minutes: anti-chromogranin A (EP38, Epitomics, 1:100), anti-cytokeratin 5/6 (QR027&QR028, Quartett, 1:300), anti-cytokeratin 7 (OV-TL 12/30, Dako, 1:1000), anti-MET (SP44, Roche/Ventana, ready-to-use), anti-p40 (SP225, Roche/Ventana, ready-to-use), anti-PD-L1 (E1L3NR, Cell Signaling, 1:100), anti-synaptophysin (27G12, Leica, 1:50), anti-TROP-2 (ENZ-ABS380, Enzo Life Sciences, 1:500), and anti-TTF1 (8G7G3/1, Zytomed, 1:100). Detection was performed using the avidin-biotin complex method with 3,3'-diaminobenzidine (DAB) as chromogen. Sections were counterstained with hematoxylin and bluing reagent (Ventana Medical Systems) for 12 minutes. For cell classification analyses, reference labels for model testing were generated by assessing immunohistochemical expression of CgA, CK5/6, CK7, p40, PD-L1, SYP, and TTF-1 using a binary classification (positive or negative). MET and TROP-2 expression was classified into four intensity categories: negative, weak, moderate, and strong. For diagnostic tumor subtyping based on immunohistochemical marker combinations, see the Tumor Subtyping via Cell Phenotyping section (Section 2.1.3). For prospective clinical validation, membranous and cytoplasmic TROP-2 expression as well as MET expression were independently assessed by five pathologists using the H-score scoring system. The percentage of positive tumor cells was estimated from 0 to 100%, and staining intensity was classified as 0 (negative), 1 (weak), 2 (moderate), or 3 (strong). H-scores were calculated as: [(percentage of weakly positive tumor cells) × 1] + [(percentage of moderately positive tumor cells) × 2] + [(percentage of strongly positive tumor cells) × 3] [80]. PD-L1 expression was assessed using the tumor proportion score (TPS), defined as the percentage of PD-L1-positive tumor cells [3, 81].

## 4.5 Image Analysis Pipeline

All image-processing modules are built on the Atlas histopathology foundation model [36, 43] and follow the Atlas H&E-TME development framework [38], to which we refer for the foundation model pre-training and for the architecture and training of the H&E tissue-profiling modules (quality control, tissue segmentation, and H&E cell phenotyping). The modules additionally developed as part of LUCAID – IHC cell phenotyping and expression scoring for PD-L1, MET, and TROP-2 – were trained within the same framework on dedicated pathologist annotations in IHC-stained whole-slide images; the IHC cell phenotyping module applies the same design as its H&E counterpart. The pipeline operates as a sequential workflow: (1) quality control identifies and masks artifact-free tissue regions, (2) tissue segmentation operates on QC-validated regions to delineate tumor boundaries, (3) cell detection and phenotyping operates within segmented tissue regions excluding blood and necrotic areas, and (4) for IHC slides, expression scoring operates on phenotyped carcinoma cells. Each module passes spatial coordinates and binary masks to downstream modules, ensuring that subsequent analyses are restricted to relevant tissue areas.

## 4.6 IHC Expression Scoring

To assess IHC expression in PD-L1, MET, and TROP-2 stained images, cells classified as carcinoma by the phenotyping model were further classified by staining intensity using a second classification model following the same architectural design. For MET and TROP-2, cells were classified into four intensity categories (negative, weak, moderate, strong) based on the membranous staining; for TROP-2, cytoplasmic staining was additionally evaluated using the same intensity categories. For PD-L1, cells were classified into two categories (negative, positive) for membranous staining. Slide-level scores (H-scores or tumor proportion scores) were then calculated by aggregating cell-level predictions across all classified tumor cells.

## 4.7 Agentic Orchestration and Report Generation

Each diagnostic module was exposed as an independently callable tool for LLM-based agent models via the Model Context Protocol (MCP), together forming the LUCAID toolbox. Tool orchestration was performed by Claude Opus 4.8.

For a given case, the agent resolved the case identifier and invoked the corresponding tools to retrieve their outputs. Every tool returned the pre-computed, validated readouts of its module, so that no quantitative value in the report was produced by the language model itself. The retrieved outputs were assembled into a report of predefined sections, each corresponding to one module. Text in the report content was generated in two stages. First, for every section, the model produced a concise clinical interpretation of each individual readout value together with a short section-level summary, conditioned on the module values and on established clinical decision thresholds. The model was constrained to restate only the provided quantities and not to introduce values of its own. Second, a separate call generated the overall assessment, synthesizing the assembled section content, the readouts, their per-readout interpretations, and the section summaries, into a case-level diagnosis and a concise account of the resulting treatment considerations. This synthesis step was explicitly restricted to information already present in the sections, so that the overall assessment condensed rather than extended the module-derived findings.

Quantitative readouts were contextualized according to their type. Diagnostic and therapeutic markers were interpreted against their established clinical cut-offs [3, 82]. Readouts that lack a single accepted threshold, such as the tissue-composition, tumor microenvironment, and immune-infiltration metrics, were instead contextualized against a reference distribution derived from an independent cohort of lung cancer cases. For each such metric, a value was placed within the reference distribution and, where a direction of clinical favorability could be defined, assigned a discrete favorability level from the corresponding cohort quantiles (low, mid, high). This placement was computed deterministically from the reference cohort rather than by the language model, and was additionally supplied to the interpretation step so that the generated text remained consistent with it. We have included the prompt that the agents use to fill the report template in Appendix C.

## 4.8 Molecular Analysis

Molecular analysis was performed as part of the routine diagnostic work-flow. For molecular profiling, tumor-rich regions were identified and annotated by pathologists using light microscopy (Olympus BX46), with tumor cellularity assessment prior to DNA extraction. Five to twenty serial Formalin-fixed, paraffin-embedded (FFPE) sections (5 $\mu$m thickness) were prepared, and DNA extraction was performed semi-automatically using the Maxwell RSC FFPE Plus DNA Purification Kit (Custom, Promega)

according to the manufacturer's protocol. DNA concentration was quantified using the Qubit HS DNA assay (Thermo Fisher Scientific).

Sequencing libraries were generated using the AmpliSeq for Illumina Cancer Hotspot nNGM Panel v3 (Illumina) with 80 ng genomic DNA input per sample. The panel covers hotspot regions of 53 cancer-associated genes: *AKT1, ALK, APC, ATM, BAP1, BRAF, BRCA1, BRCA2, CHEK2, CTNNB1, CUL3, DPYD, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FGFR4, GNA11, GNAQ, GNAS, HRAS, IDH1, IDH2, JAK2, KEAP1, KIT, KRAS, MAP2K1, MEN1, MET, MLH1, MSH2, MSH6, NF1, NFE2L2, NRAS, NTRK1, NTRK2, NTRK3, PALB2, PDGFRA, PDGFRB, PIK3CA, PMS2, PTEN, RB1, RET, ROS1, SMARCA4, STK11, TERT*, and *TP53*. Target regions were amplified by PCR using a Biometra TOne thermal cycler (Analytik Jena, Jena, Germany), followed by next-generation sequencing on an NextSeq Sequencing System (Illumina). Sequencing reads were aligned to the human reference genome (hg19), and variant calling was performed using SEQUENCE Pilot Software version 5.4.0 (JSI Medical Systems GmbH, Ettenheim, Germany). Variants were filtered using a minimum allele frequency threshold of 5%, and pathogenic or likely pathogenic variants were retained for downstream analysis.

## 4.9 Expert-Panel Adjudicated Reference Standard

To establish the reference standard for the prospective clinical validation, all five participating thoracic pathologists re-evaluated all 70 cases after a washout period of at least three months following completion of the initial assessment of the final case. Each case was jointly reviewed by the expert panel, and a final task-specific score and clinical classification were reached by discussion.

Expert-panel adjudication was chosen over simple majority voting because joint expert review can reduce the influence of individual observer variability and allows discrepant assessments to be resolved through case-level re-evaluation [83, 84]. Majority voting, by contrast, may preserve systematic estimation biases when several observers assess a feature in the same direction. This is particularly relevant for tumor cellularity, for which visual assessment shows substantial interobserver variability and has been reported to overestimate tumor content in a relevant proportion of cases [53, 54, 85]. This observation was also consistent with our data. Within the reference-standard cohort itself, the adjudicated tumor cellularity was systematically lower than the mean of the individual panel members (median $-10$ ppt; lower in 86% of cases; $p < 10^{-9}$, Wilcoxon signed-rank test), and four of the five raters scored predominantly above the adjudicated value, indicating a shared, same-direction tendency among the individual raters that joint review resolved but a simple majority vote would have retained.

During adjudication, the histopathological slides were reviewed together with LUCAID-derived heatmaps. These overlays are purely qualitative spatial visualizations of tissue and cell distributions; the panel was not shown the corresponding numeric module readouts (for example, the model-derived tumor cellularity percentage, PD-L1 TPS, or MET and TROP-2 H-scores) or LUCAID's categorical classifications. The overlays therefore facilitated visual inspection by displaying tissue and cell distributions with greater contrast than the underlying histopathological image, thereby making semiquantitative features easier to assess visually, without disclosing the model's predicted values and thus limiting the risk of anchoring the adjudicated scores to LUCAID's output. Previous studies have likewise shown that AI-supported visualization or quantitative assistance can improve the consistency of tumor cellularity assessment by pathologists [86–88]. Review of the heatmaps also allowed the panel to assess the plausibility of the model-derived spatial classifications in their histopathological context, providing an additional quality-control layer before their use as visual support during adjudication. The heatmaps served as an adjunct to pathological review rather than as an independent reference; final scores and clinical classifications were determined by the expert panel after joint evaluation and discussion.

The resulting adjudicated assessments constituted the expert-panel adjudicated reference standard used for all clinical-validation analyses.

## 4.10 Clinical Action Concordance

To evaluate the agreement of LUCAID-derived continuous diagnostic scores with clinically relevant decision categories, we defined a clinical action concordance index (CAC). Continuous predictions from LUCAID and pathologist assessments were converted into clinically relevant categorical classifications using established thresholds and compared with the corresponding reference-standard classifications (see Immunohistochemical Analysis section for further details). The following clinical decision categories were applied: molecular testing adequacy based on tumor cellularity ($< 10\%$ versus $\geq 10\%$), PD-L1 expression based on tumor proportion score (TPS; $< 1\%$, 1–49% and $\geq 50\%$), and MET as well as membranous and cytoplasmic TROP-2 expression based on H-score categories (negative/weak [0–100); moderate [100–200) and high [200–300]). CAC was calculated as the proportion of cases in which the categorical assessment of each rater matched the corresponding reference-standard classification. In this analysis the LUCAID rater denotes the readouts of the individual diagnostic modules rather than the agent's free-text output, which was evaluated separately for groundedness and interpretation (see the Report Generation results). Agreement was evaluated independently for each clinical action category.

## 4.11 Pathologic Regression Grading After Neoadjuvant Therapy

Pathologic response to neoadjuvant therapy was assessed in 140 resected non-small cell lung cancer cases treated with neoadjuvant chemotherapy or chemoimmunotherapy at the Evangelische Lungenklinik Berlin-Buch between 2020 and 2026 (Figure 6d–f), following the IASLC multidisciplinary recommendations for pathologic assessment of lung cancer resection specimens following neoadjuvant therapy [59]. Two pathologists jointly reviewed all H&E-stained slides of the tumor bed and estimated the percentages of (1) viable tumor, (2) necrosis, and (3) stroma, the latter comprising both fibrosis and inflammation, with the three components summing to 100% of the tumor bed. In accordance with these recommendations, each component was recorded in 10% increments, except for amounts of 5% or less, which were recorded in single-percentage steps. For automated assessment, LUCAID quantified the same three components as area fractions of the segmented carcinoma, stromal, and necrotic compartments within the tumor region derived from the tissue segmentation module. Agreement between model-derived and pathologist estimates was assessed per component using Pearson and Spearman correlation coefficients (two-tailed, $\alpha = 0.05$), and quantitative deviation was summarized as the mean absolute error of the model against the joint pathologist assessment.

## 4.12 Spatial Feature and Survival Analysis

Spatial tumor-microenvironment features were computed per tissue sample from AI-phenotyped cells, restricted to the seven analysis phenotypes (carcinoma, endothelial, fibroblast, granulocyte, lymphocyte, macrophage, and plasma cells) and averaged across tissue samples to the patient level. Each feature was related to overall survival in its own Cox proportional-hazards model together with UICC stage (stage-adjusted, not co-fitted), reporting the hazard ratio, 95% confidence interval, and log-rank p with Benjamini–Hochberg false-discovery rate control across the panel, and was dichotomized at the median – or, for bimodal or strongly skewed markers, at a Gaussian-mixture threshold placed at the trough between the two populations (rule and cut per feature in Supplementary Table 3). The novel spatial features comprised three families: niche fractions, obtained by building a Delaunay neighbor graph over the cells, partitioning it into spatial communities by compartment-aware Leiden community

detection, and typing each community by its cell-type composition — a community being lymphoid (tertiary-lymphoid-structure–like) when lymphocytes plus plasma cells reached $\geq 40\%$ of its cells—with the feature value defined as the fraction of analyzed cells assigned to that niche; adjacency fractions, the proportion of neighbor-graph edges directly connecting a given cell-type pair (e.g., carcinoma–plasma cell); and cell–cell distances, written A–B, the median over the A cells of the distance to the nearest anchoring B cell.

## 4.13 Statistical Analysis

Module performance for quality control, tissue segmentation, and cell phenotyping was evaluated with F1 scores, reported per class and as macro averages, comparing predictions to pathologist annotations at the pixel level (segmentation) or cell level (classification); IHC expression-scoring models were evaluated analogously with per-category and macro-averaged F1 on the annotated cells of each slide.

Continuous predictions were compared with their reference standard (expert-panel consensus or, for tumor cellularity, *KRAS* VAF) using Pearson and Spearman correlation coefficients (two-tailed, $\alpha = 0.05$) and mean absolute error (MAE) with 95% confidence intervals from 1,000 bootstrap iterations. Slides with fewer than 100 tumor cells were excluded from percentage-based analyses. All analyses used Python (scipy.stats, numpy).

## 4.14 Ethical Approval and Consent

Ethical approval for this study was granted by the Ethics Committee of Charité – Universitätsmedizin Berlin (EA4/082/22), and all procedures adhered to the ethical principles for medical research of the Declaration of Helsinki. All patients provided written informed consent for the scientific use of their archived tissue and associated clinical data. Information on race, ethnicity and socioeconomic status was not collected as part of routine diagnostic documentation and was therefore unavailable for analysis.

## Acknowledgements

Marie-Lisa Eich is a participant in the BIH Charité Digital Clinician Scientist Program funded by the Charité – University Medical Center Berlin and the Berlin Institute of Health at Charité (BIH). Marie-Lisa Eich and Simon Schallenberg were co-funded by the European Union and the Investment Bank Berlin PROFIT grant (grant: 10191964). Mihnea P. Dragomir is a participant in the BIH Charité Clinician Scientist Program funded by the Charité – University Medical Center Berlin, and the Berlin Institute of Health at Charité (BIH). This work was partly funded by the German Ministry for Education and Research (under refs 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18056A, 01IS18025A, 13GW0744D, and BIFOLD25B) and by DFG. Furthermore, Klaus-Robert Müller was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program, Korea University) and grant funded by the Korea government (MSIT) (No. RS-2024-00457882, AI Research Hub Project).

## Competing interests

F.K., M.A., and K.R.M. are co-founders of Aignostics. F.K. is lead medical advisor and board member of Aignostics. K.R.M. is lead technical advisor to Aignostics. M.A. is CTO of Aignostics. S.S. are part-time employees at Aignostics. DH is a member of the scientific advisory board at Aignostics. The other authors declare no conflict of interest.

## References

[1] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. ISSN 1542-4863. doi: 10.3322/caac.21834. URL https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21834.

[2] Nadia Howlader, Gonçalo Forjaz, Meghan J. Mooradian, Rafael Meza, Chung Yin Kong, Kathleen A. Cronin, Angela B. Mariotto, Douglas R. Lowy, and Eric J. Feuer. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *New England Journal of Medicine*, 383(7): 640–649, August 2020. ISSN 0028-4793. doi: 10.1056/NEJMoa1916623. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1916623. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1916623.

[3] Martin Reck, Delvys Rodríguez-Abreu, Andrew G. Robinson, Rina Hui, Tibor Csőszi, Andrea Fülöp, Maya Gottfried, Nir Peled, Ali Tafreshi, Sinead Cuffe, Mary O'Brien, Suman Rao, Katsuyuki Hotta, Melanie A. Leiby, Gregory M. Lubiniecki, Yue Shentu, Reshma Rangwala, and Julie R. Brahmer. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 375(19):1823–1833, November 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1606774. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1606774. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1606774.

[4] Yi-Long Wu, Masahiro Tsuboi, Jie He, Thomas John, Christian Grohe, Margarita Majem, Jonathan W. Goldman, Konstantin Laktionov, Sang-We Kim, Terufumi Kato, Huu-Vinh Vu, Shun Lu, Kye-Young Lee, Charuwan Akewanlop, Chong-Jen Yu, Filippo de Marinis,

Laura Bonanno, Manuel Domine, Frances A. Shepherd, Lingmin Zeng, Rachel Hodge, Ajlan Atasoy, Yuri Rukazenkov, and Roy S. Herbst. Osimertinib in Resected EGFR-Mutated Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 383(18):1711–1723, October 2020. ISSN 0028-4793. doi: 10.1056/NEJMoa2027071. URL https://www.nejm.org/doi/full/10.1056/NEJMoa2027071. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2027071.

[5] Enriqueta Felip, Nasser Altorki, Caicun Zhou, Tibor Csőszi, Ihor Vynnychenko, Oleksandr Goloborodko, Alexander Luft, Andrey Akopov, Alex Martinez-Marti, Hirotsugu Kenmotsu, Yuh-Min Chen, Antonio Chella, Shunichi Sugawara, David Voong, Fan Wu, Jing Yi, Yu Deng, Mark McCleland, Elizabeth Bennett, Barbara Gitlitz, and Heather Wakelee. Adjuvant atezolizumab after adjuvant chemotherapy in resected stage IB–IIIA non-small-cell lung cancer (IMpower010): a randomised, multicentre, open-label, phase 3 trial. *The Lancet*, 398(10308):1344–1357, October 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(21)02098-5. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02098-5/fulltext. Publisher: Elsevier.

[6] Anika Kästner, Anna Kron, Neeltje van den Berg, Kilson Moon, Matthias Scheffler, Gerhard Schillinger, Natalie Pelusi, Nils Hartmann, Damian Tobias Rieke, Susann Stephan-Falkenau, Martin Schuler, Martin Wermke, Wilko Weichert, Frederick Klauschen, Florian Haller, Horst-Dieter Hummel, Martin Sebastian, Stefan Gattenlöhner, Carsten Bokemeyer, Irene Esposito, Florian Jakobs, Christof von Kalle, Reinhard Büttner, Jürgen Wolf, and Wolfgang Hoffmann. Evaluation of the effectiveness of a nationwide precision medicine program for patients with advanced non-small cell lung cancer in Germany: a historical cohort analysis. *The Lancet Regional Health – Europe*, 36, January 2024. ISSN 2666-7762. doi: 10.1016/j.lanepe.2023.100788. URL https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(23)00207-7/fulltext. Publisher: Elsevier.

[7] A. Passaro, N. Leighl, F. Blackhall, S. Popat, K. Kerr, M. J. Ahn, M. E. Arcila, O. Arrieta, D. Planchard, F. de Marinis, A. M. Dingemans, R. Dziadziuszko, C. Faivre-Finn, J. Feldman, E. Felip, G. Curigliano, R. Herbst, P. A. Jänne, T. John, T. Mitsudomi, T. Mok, N. Normanno, L. Paz-Ares, S. Ramalingam, L. Sequist, J. Vansteenkiste, I. I. Wistuba, J. Wolf, Y. L. Wu, S. R. Yang, J. C. H. Yang, Y. Yatabe, G. Pentheroudakis, and S. Peters. ESMO expert consensus statements on the management of *EGFR* mutant non-small-cell lung cancer. *Annals of Oncology*, 33(5):466–487, May 2022. ISSN 0923-7534. doi: 10.1016/j.annonc.2022.02.003. URL https://www.sciencedirect.com/science/article/pii/S0923753422001120.

[8] Nasser H. Hanna, Andrew G. Robinson, Sarah Temin, Sherman Baker, Julie R. Brahmer, Peter M. Ellis, Laurie E. Gaspar, Rami Y. Haddad, Paul J. Hesketh, Dharamvir Jain, Ishmael Jaiyesimi, David H. Johnson, Natasha B. Leighl, Pamela R. Moffitt, Tanyanika Phillips, Gregory J. Riely, Rafael Rosell, Joan H. Schiller, Bryan J. Schneider, Navneet Singh, David R. Spigel, Joan Tashbar, and Gregory Masters. Therapy for Stage IV Non-Small-Cell Lung Cancer With Driver Alterations: ASCO and OH (CCO) Joint Guideline Update. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 39(9):1040–1091, March 2021. ISSN 1527-7755. doi: 10.1200/JCO.20.03570.

[9] A. Zer, M.-J. Ahn, F. Barlesi, L. Bubendorf, D. De Ruysscher, P. Garrido, O. Gautschi, L. E. Hendriks, P. A. Jänne, K. M. Kerr, C. Mascaux, T. Mitsudomi, S. Peters, C. Rolfo, A. Sacher, S. Senan, P. Ugalde, and N. B. Leighl. Early and locally advanced non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up⋆. *Annals of Oncology*, 36 (11):1245–1262, November 2025. ISSN 0923-7534, 1569-8041. doi: 10.1016/j.annonc.2025.08.003. URL https://www.annalsofoncology.org/article/S0923-7534(25)00923-8/fulltext.

[10] L. E. Hendriks, K. M. Kerr, J. Menis, T. S. Mok, U. Nestle, A. Passaro, S. Peters, D. Planchard, E. F. Smit, B. J. Solomon, G. Veronesi, M. Reck, and ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. Oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 34(4):339–357, April 2023. ISSN 1569-8041. doi: 10.1016/j.annonc.2022.12.009.

[11] D. Ross Camidge, Jair Bar, Hidehito Horinouchi, Jonathan Goldman, Fedor Moiseenko, Elena Filippova, Irfan Cicin, Tudor Ciuleanu, Nathalie Daaboul, Chunling Liu, Penelope Bradbury, Mor Moskovitz, Nuran Katgi, Pascale Tomasini, Alona Zer, Nicolas Girard, Kristof Cuppens, Ji-Youn Han, Shang-Yin Wu, Shobhit Baijal, Aaron S. Mansfield, Chih-Hsi Kuo, Kazumi Nishino, Se-Hoon Lee, David Planchard, Christina Baik, Martha Li, Peter Ansell, Summer Xia, Ellen Bolotin, Jim Looman, Christine Ratajczak, and Shun Lu. Telisotuzumab Vedotin Monotherapy in Patients With Previously Treated c-Met Protein-Overexpressing Advanced Nonsquamous EGFR-Wildtype Non-Small Cell Lung Cancer in the Phase II LUMINOSITY Trial. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 42(25):3000–3011, September 2024. ISSN 1527-7755. doi: 10.1200/JCO.24.00720.

[12] Leah Lawrence. FDA approves second-line TROP2-directed antibody–drug conjugate Dato-DXd for EGFR-mutated NSCLC. *Cancer*, 131(21):e70090, 2025. ISSN 1097-0142. doi: 10.1002/cncr.70090. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.70090. _eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.70090.

[13] Myung-Ju Ahn, Kentaro Tanaka, Luis Paz-Ares, Robin Cornelissen, Nicolas Girard, Elvire Pons-Tostivint, David Vicente Baz, Shunichi Sugawara, Manuel Cobo, Maurice Pérol, Céline Mascaux, Elena Poddubskaya, Satoru Kitazono, Hidetoshi Hayashi, Min Hee Hong, Enriqueta Felip, Richard Hall, Oscar Juan-Vidal, Daniel Brungs, Shun Lu, Marina Garassino, Michael Chargualaf, Yong Zhang, Paul Howarth, Deise Uema, Aaron Lisberg, Jacob Sands, and TROPION-Lung01 Trial Investigators. Datopotamab Deruxtecan Versus Docetaxel for Previously Treated Advanced or Metastatic Non-Small Cell Lung Cancer: The Randomized, Open-Label Phase III TROPION-Lung01 Study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 43(3):260–272, January 2025. ISSN 1527-7755. doi: 10.1200/JCO-24-01544.

[14] Eoghan R. Malone, Marc Oliva, Peter J. B. Sabatini, Tracy L. Stockley, and Lillian L. Siu. Molecular profiling for precision cancer therapies. *Genome Medicine*, 12(1):8, January 2020. ISSN 1756-994X. doi: 10.1186/s13073-019-0703-1.

[15] Elizabeth Walsh and Nicolas M. Orsi. The current troubled state of the global pathology workforce: a concise review. *Diagnostic Pathology*, 19(1):163, December 2024. ISSN 1746-1596. doi: 10.1186/s13000-024-01590-2.

[16] Kris Lami, Andrey Bychkov, Keitaro Matsumoto, Richard Attanoos, Sabina Berezowska, Luka Brcic, Alberto Cavazza, John C. English, Alexandre Todorovic Fabro, Kaori Ishida, Yukio Kashima, Brandon T. Larsen, Alberto M. Marchevsky, Takuro Miyazaki, Shimpei Morimoto, Anja C. Roden, Frank Schneider, Mano Soshi, Maxwell L. Smith, Kazuhiro Tabata, Angela M. Takano, Kei Tanaka, Tomonori Tanaka, Tomoshi Tsuchiya, Takeshi Nagayasu, and Junya Fukuoka. Overcoming the Interobserver Variability in Lung Adenocarcinoma Subtyping: A Clustering Approach to Establish a Ground Truth for Downstream Applications. *Archives of Pathology & Laboratory Medicine*, 147(8):885–895, August 2023. ISSN 1543-2165. doi: 10.5858/arpa.2022-0051-OA.

[17] Marie E. Robert, Josef Rüschoff, Bharat Jasani, Rondell P. Graham, Sunil S. Badve, Manuel Rodriguez-Justo, Liudmila L. Kodach, Amitabh Srivastava, Hanlin L. Wang, Laura H. Tang,

Giancarlo Troncone, Federico Rojo, Benjamin J. Van Treeck, James Pratt, Iryna Shnitsar, George Kumar, Maria Karasarides, and Robert A. Anders. High Interobserver Variability Among Pathologists Using Combined Positive Score to Evaluate PD-L1 Expression in Gastric, Gastroesophageal Junction, and Esophageal Adenocarcinoma. *Modern Pathology*, 36 (5), May 2023. ISSN 0893-3952, 1530-0285. doi: 10.1016/j.modpat.2023.100154. URL https://www.modernpathology.org/article/S0893-3952(23)00059-5/fulltext.

[18] Zoya Volynskaya, Ozgur Mete, Sara Pakbaz, Doaa Al-Ghamdi, and Sylvia L. Asa. Ki67 Quantitative Interpretation: Insights using Image Analysis. *Journal of Pathology Informatics*, 10:8, 2019. ISSN 2229-5089. doi: 10.4103/jpi.jpi_76_18.

[19] Sehhoon Park, Chan-Young Ock, Hyojin Kim, Sergio Pereira, Seonwook Park, Minuk Ma, Sangjoon Choi, Seokhwi Kim, Seunghwan Shin, Brian Jaehong Aum, Kyunghyun Paeng, Donggeun Yoo, Hongui Cha, Sunyoung Park, Koung Jin Suh, Hyun Ae Jung, Se Hyun Kim, Yu Jung Kim, Jong-Mu Sun, Jin-Haeng Chung, Jin Seok Ahn, Myung-Ju Ahn, Jong Seok Lee, Keunchil Park, Sang Yong Song, Yung-Jue Bang, Yoon-La Choi, Tony S. Mok, and Se-Hoon Lee. Artificial Intelligence-Powered Spatial Analysis of Tumor-Infiltrating Lymphocytes as Complementary Biomarker for Immune Checkpoint Inhibition in Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 40(17):1916–1928, June 2022. ISSN 1527-7755. doi: 10.1200/JCO.21.02010.

[20] Qin Yan, Shuai Li, Lang He, and Nianyong Chen. Prognostic implications of tumor-infiltrating lymphocytes in non-small cell lung cancer: a systematic review and meta-analysis. *Frontiers in Immunology*, 15:1476365, September 2024. ISSN 1664-3224. doi: 10.3389/fimmu.2024.1476365. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11449740/.

[21] Bonan Chen, Xiaohong Zheng, Jialin Wu, Guoming Chen, Jun Yu, Yi Xu, William K. K. Wu, Gary M. K. Tse, Ka Fai To, and Wei Kang. Antibody-drug conjugates in cancer therapy: current landscape, challenges, and future directions. *Molecular Cancer*, 24(1):279, November 2025. ISSN 1476-4598. doi: 10.1186/s12943-025-02489-2.

[22] Markus Moehler, Harry H. Yoon, Daniel-Christoph Wagner, Silu Yang, Jingwen Shi, Yun Zhang, Han Hu, Christopher La Placa, Yanyan Peng, Wenting Du, Adrienne McCampbell, Wenjie Xu, Zhirong Shen, Hui Xu, Ruiqi Huang, and Ken Kato. Concordance Between the PD-L1 Tumor Area Positivity Score and Combined Positive Score for Gastric or Esophageal Cancers Treated With Tislelizumab. *Modern Pathology*, 38(9), September 2025. ISSN 0893-3952, 1530-0285. doi: 10.1016/j.modpat.2025.100793. URL https://www.modernpathology.org/article/S0893-3952(25)00089-4/fulltext.

[23] Louis Fehrenbacher, Alexander Spira, Marcus Ballinger, Marcin Kowanetz, Johan Vansteenkiste, Julien Mazieres, Keunchil Park, David Smith, Angel Artal-Cortes, Conrad Lewanski, Fadi Braiteh, Daniel Waterkamp, Pei He, Wei Zou, Daniel S. Chen, Jing Yi, Alan Sandler, and Achim Rittmeyer. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *The Lancet*, 387(10030):1837–1846, April 2016. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(16)00587-0. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)00587-0/abstract.

[24] Edward B. Garon, Naiyer A. Rizvi, Rina Hui, Natasha Leighl, Ani S. Balmanoukian, Joseph Paul Eder, Amita Patnaik, Charu Aggarwal, Matthew Gubens, Leora Horn, Enric Carcereny, Myung-Ju Ahn, Enriqueta Felip, Jong-Seok Lee, Matthew D. Hellmann, Omid Hamid, Jonathan W. Goldman, Jean-Charles Soria, Marisa Dolled-Filhart, Ruth Z. Rutledge, Jin Zhang, Jared K.

Lunceford, Reshma Rangwala, Gregory M. Lubiniecki, Charlotte Roach, Kenneth Emancipator, and Leena Gandhi. Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 372(21):2018–2028, May 2015. ISSN 0028-4793. doi: 10.1056/NEJMoa1501824. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1501824. _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1501824.

[25] Misty M. Attwood, Doriano Fabbro, Aleksandr V. Sokolov, Stefan Knapp, and Helgi B. Schiöth. Trends in kinase drug discovery: targets, indications and inhibitor design. *Nature Reviews. Drug Discovery*, 20(11):839–861, November 2021. ISSN 1474-1784. doi: 10.1038/s41573-021-00252-y.

[26] Sunhee Chang, Hyung Kyu Park, Yoon-La Choi, and Se Jin Jang. Interobserver Reproducibility of PD-L1 Biomarker in Non-small Cell Lung Cancer: A Multi-Institutional Study by 27 Pathologists. *Journal of Pathology and Translational Medicine*, 53(6):347–353, November 2019. ISSN 2383-7837. doi: 10.4132/jptm.2019.09.29.

[27] Rogier Butter, Liesbeth M. Hondelink, Lisette van Elswijk, Johannes L. G. Blaauwgeers, Elisabeth Bloemena, Rieneke Britstra, Nicole Bulkmans, Anna Lena van Gulik, Kim Monkhorst, Mathilda J. de Rooij, Ivana Slavujevic-Letic, Vincent T. H. B. M. Smit, Ernst-Jan M. Speel, Erik Thunnissen, Jan H. von der Thüsen, Wim Timens, Marc J. van de Vijver, David C. Y. Yick, Aeilko H. Zwinderman, Danielle Cohen, Nils A. 't Hart, and Teodora Radonic. The impact of a pathologist's personality on the interobserver variability and diagnostic accuracy of predictive PD-L1 immunohistochemistry in lung cancer. *Lung Cancer*, 166:143–149, April 2022. ISSN 1872-8332. doi: 10.1016/j.lungcan.2022.03.002. Place: Amsterdam, Netherlands.

[28] Mieke R. Van Bockstal, Marie-Caroline Depelsemaeker, Lina Daoud, Quitterie Fontanges, Aline Francois, Yves Guiot, Anne-France Dekairelle, Dominique Dubois, Cédric Van Marcke, Eléonore Longton, Francois P. Duhoux, Hilde Vernaeve, Martine Berlière, Giuseppe Floris, and Christine Galant. Evaluation of trophoblast cell surface antigen-2 (TROP2) protein expression in chemotherapy-resistant and metastatic breast carcinomas. *Pathology, Research and Practice*, 264:155724, December 2024. ISSN 1618-0631. doi: 10.1016/j.prp.2024.155724.

[29] Christophe Bontoux, Véronique Hofman, Emmanuel Chamorey, Renaud Schiappa, Sandra Lassalle, Elodie Long-Mira, Katia Zahaf, Salomé Lalvée, Julien Fayada, Christelle Bonnetaud, Samantha Goffinet, Marius Ilié, and Paul Hofman. Reproducibility of c-Met Immunohistochemical Scoring (Clone SP44) for Non-Small Cell Lung Cancer Using Conventional Light Microscopy and Whole Slide Imaging. *The American Journal of Surgical Pathology*, 48(9):1072–1081, September 2024. ISSN 1532-0979. doi: 10.1097/PAS.0000000000002274.

[30] Simon Schallenberg, Gabriel Dernbach, Sharon Ruane, Philipp Jurmeister, Cornelius Böhm, Kai Standvoss, Sandip Ghosh, Marco Frentsch, Mihnea P. Dragomir, Philipp G. Keyl, Corinna Friedrich, Il-Kang Na, Sabine Merkelbach-Bruse, Alexander Quaas, Nikolaj Frost, Kyrill Boschung, Winfried Randerath, Georg Schlachtenberger, Matthias Heldwein, Ulrich Keilholz, Khosro Hekmat, Jens-Carsten Rückert, Reinhard Büttner, Angela Vasaturo, David Horst, Lukas Ruff, Maximilian Alber, Klaus-Robert Müller, and Frederick Klauschen. AI-powered spatial cell phenomics enhances risk stratification in non-small cell lung cancer. *Nature Communications*, 16(1):9701, November 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-65783-z. URL https://www.nature.com/articles/s41467-025-65783-z. Publisher: Nature Publishing Group.

[31] Edwin Roger Parra, Jiexin Zhang, Mei Jiang, Auriole Tamegnon, Renganayaki Krishna Pandurengan, Carmen Behrens, Luisa Solis, Cara Haymaker, John Victor Heymach, Cesar Moran, Jack J. Lee, Don Gibbons, and Ignacio Ivan Wistuba. Immune cellular patterns of distribution affect outcomes of patients with non-small cell lung cancer. *Nature Communications*, 14(1):2364, April 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37905-y.

[32] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, October 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03141-0. URL https://www.nature.com/articles/s41591-024-03141-0.

[33] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07441-w. URL https://www.nature.com/articles/s41586-024-07441-w.

[34] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3. Publisher: Nature Publishing Group.

[35] Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Timo Milbich, Stephan Tietz, Simon Schallenberg, Gabriel Dernbach, Andreas Kunft, Simon Heinke, Marie-Lisa Eich, Julika Ribbat-Idel, Rosemarie Krupar, Philipp Anders, Niklas Prenißl, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Maximilian Alber. RudolfV: A Foundation Model by Pathologists for Pathologists, June 2024. URL http://arxiv.org/abs/2401.04079. arXiv:2401.04079 [eess.IV].

[36] Maximilian Alber, Stephan Tietz, Jonas Dippel, Timo Milbich, Timothée Lesort, Panos Korfiatis, Moritz Krügener, Beatriz Perez Cancer, Neelay Shah, Alexander Möllers, Philipp Seegerer, Alexandra Carpen-Amarie, Kai Standvoss, Gabriel Dernbach, Edwin de Jong, Simon Schallenberg, Andreas Kunft, Helmut Hoffer von Ankershoffen, Gavin Schaeferle, Patrick Duffy, Matt Redlon, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Andrew Norgan. Atlas: A Novel Pathology Foundation Model by Mayo Clinic, Charité, and Aignostics, January 2025. URL http://arxiv.org/abs/2501.05409. arXiv:2501.05409 [cs].

[37] Gabriele Campanella, Neeraj Kumar, Swaraj Nanda, Siddharth Singi, Eugene Fluder, Ricky Kwan, Silke Muehlstedt, Nicole Pfarr, Peter J. Schüffler, Ida Häggström, Noora Neittaanmäki, Levent M. Akyürek, Alina Basnet, Tamara Jamaspishvili, Michel R. Nasr, Matthew M. Croken, Fred R. Hirsch, Arielle Elkrief, Helena Yu, Orly Ardon, Gregory M. Goldgof, Meera Hameed, Jane Houldsworth, Maria Arcila, Thomas J. Fuchs, and Chad Vanderbilt. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine*, 31 (9):3002–3010, September 2025. ISSN 1546-170X. doi: 10.1038/s41591-025-03780-x. URL https://www.nature.com/articles/s41591-025-03780-x. Publisher: Nature Publishing Group.

[38] Kai Standvoss, Miriam Hägele, Rosemarie Krupar, Julika Ribbat-Idel, Jennifer Altschüler, Gerrit Erdmann, Hans Pinckaers, Evelyn Ramberger, Madleen Drinkwitz, Ádám Nárai, Alexander Möllers,

Katja Lingelbach, Sebastian Kons, Lukas Hönig, Recepcan Adigüzel, Joana Baião, Alberto Megina Gonzalo, Marius Teodorescu, Marie-Lisa Eich, Paolo Chetta, Shakil Merchant, Verena Aumiller, Simon Schallenberg, Andrew Norgan, Klaus-Robert Müller, Lukas Ruff, Maximilian Alber, and Frederick Klauschen. Atlas H&E-TME: Scalable AI-Based Tissue Profiling at Expert Pathologist-Level Accuracy, July 2026. URL http://arxiv.org/abs/2606.12346. arXiv:2606.12346 [cs.CV].

[39] Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer*, 6(8):1337–1349, August 2025. ISSN 2662-1347. doi: 10.1038/s43018-025-00991-6. URL https://www.nature.com/articles/s43018-025-00991-6.

[40] Dyke Ferber, Isabella C. Wiest, Georg Wölflein, Matthias P. Ebert, Gernot Beutel, Jan-Niklas Eckardt, Daniel Truhn, Christoph Springfeld, Dirk Jäger, and Jakob Nikolas Kather. GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. *NEJM AI*, 1(6): AIcs2300235, May 2024. doi: 10.1056/AIcs2300235. URL https://ai.nejm.org/doi/abs/10.1056/AIcs2300235.

[41] Alex J. Goodell, Simon N. Chu, Dara Rouholiman, and Larry F. Chu. Large language model agents can use tools to perform clinical calculations. *npj Digital Medicine*, 8(1):163, March 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01475-8. URL https://www.nature.com/articles/s41746-025-01475-8.

[42] Luca L. Weishaupt, Chengkuan Chen, Drew F. K. Williamson, Richard J. Chen, Guillaume Jaume, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Guillaume Jaume, Ming Y. Lu, and Faisal Mahmood. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology, March 2026. URL http://arxiv.org/abs/2506.20964. arXiv:2506.20964 [cs.CV].

[43] Maximilian Alber, Timo Milbich, Alexandra Carpen-Amarie, Stephan Tietz, Jonas Dippel, Lukas Muttenthaler, Beatriz Perez Cancer, Alessandro Benetti, Panos Korfiatis, Elias Eulig, Jérôme Lüscher, Jiasen Wu, Sayed Abid Hashimi, Gabriel Dernbach, Simon Schallenberg, Neelay Shah, Moritz Krügener, Aniruddh Jammoria, Jake Matras, Patrick Duffy, Matt Redlon, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Andrew Norgan. Atlas 2 – Foundation models for clinical deployment, July 2026. URL http://arxiv.org/abs/2601.05148. arXiv:2601.05148 [cs.CV].

[44] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. *Second Conference on Language Modeling*, 2025.

[45] Hong-Tao Yu, Yuxin Peng, Serge Belongie, and Xiu-Shen Wei. Benchmarking Large Vision-Language Models on Fine-Grained Image Tasks: A Comprehensive Evaluation. *The Fourteenth International Conference on Learning Representations*, 2026.

[46] Ivan Kukuljan, Muhammed Furkan Dasdelen, Julia Schäfer, Michele Buck, Katharina S. Götze, and Carsten Marr. Illusion of competence: vision–language models provide confident but inaccurate explanations in cytological diagnostics. *Scientific Reports*, 16(1):20526, July 2026. ISSN 2045-2322. doi: 10.1038/s41598-026-60372-6. URL https://www.nature.com/articles/s41598-026-60372-6. Publisher: Nature Publishing Group.

[47] Zongyi Chen, Yu Liang, Jie Lin, and Liansheng Wang. PathView-Bench: Can Multimodal Large Language Models Achieve Fine-grained Multiscale Understanding of Pathology Images?, July 2026. URL http://arxiv.org/abs/2607.28318. arXiv:2607.28318 [cs.AI].

[48] Zhilong Weng, Alexander Seper, Alexey Pryalukhin, Fabian Mairinger, Claudia Wickenhauser, Marcus Bauer, Lennert Glamann, Hendrik Bläker, Thomas Lingscheidt, Wolfgang Hulla, Danny Jonigk, Simon Schallenberg, Andrey Bychkov, Junya Fukuoka, Martin Braun, Birgid Schömig-Markiefka, Sebastian Klein, Andreas Thiel, Katarzyna Bozek, George J. Netto, Alexander Quaas, Reinhard Büttner, and Yuri Tolkach. GrandQC: A comprehensive solution to quality control problem in digital pathology. *Nature Communications*, 15(1):10685, December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54769-y. URL https://www.nature.com/articles/s41467-024-54769-y.

[49] Falah Jabar, Lill-Tove Rasmussen Busund, Biagio Ricciuti, Masoud Tafavvoghi, Thomas K Kilvaer, J Pinato, Mette Pøhl, Sigve Andersen, Tom Donnem, David J Kwiatkowski, and Mehrdad Rakaee. Fully Automatic Content-Aware Tiling Pipeline for Pathology Whole Slide Images.

[50] Abhijeet Patil, Harsh Diwakar, Jay Sawant, Nikhil Cherian Kurian, Subhash Yadav, Swapnil Rane, Tripti Bameta, and Amit Sethi. Efficient quality control of whole slide pathology images with human-in-the-loop training. *Journal of Pathology Informatics*, 14:100306, January 2023. ISSN 2153-3539. doi: 10.1016/j.jpi.2023.100306. URL https://www.sciencedirect.com/science/article/pii/S2153353923001207.

[51] Neel Kanwal, Farbod Khoraminia, Umay Kiraz, Andres Mosquera-Zamudio, Carlos Monteagudo, Emiel A. M. Janssen, Tahlita C. M. Zuiverloon, Chunmig Rong, and Kjersti Engan. Equipping Computational Pathology Systems with Artifact Processing Pipelines: A Showcase for Computation and Performance Trade-offs, May 2024. URL http://arxiv.org/abs/2403.07743. arXiv:2403.07743 [eess.IV].

[52] WHO Classification of Tumours Editorial Board. *Thoracic tumours*, volume 5 of *WHO classification of tumours series*. International Agency for Research on Cancer, Lyon (France), 5 edition, 2021. ISBN 978-92-832-4506-3. URL https://tumourclassification.iarc.who.int/chapters/35.

[53] Alexander J. J. Smits, J. Alain Kummer, Peter C. de Bruin, Mijke Bol, Jan G. van den Tweel, Kees A. Seldenrijk, Stefan M. Willems, G. Johan A. Offerhaus, Roel A. de Weger, Paul J. van Diest, and Aryan Vink. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 27(2):168–174, February 2014. ISSN 1530-0285. doi: 10.1038/modpathol.2013.134.

[54] Masashi Mikubo, Katsutoshi Seto, Atsuko Kitamura, Masato Nakaguro, Yukinori Hattori, Nagako Maeda, Tatsuhiko Miyazaki, Kazuko Watanabe, Hideki Murakami, Tetsuya Tsukamoto, Tetsuya Yamada, Shiro Fujita, Katsuhiro Masago, Shakti Ramkissoon, Jeffrey S. Ross, Julia Elvin, and Yasushi Yatabe. Calculating the Tumor Nuclei Content for Comprehensive Cancer Panel Testing. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 15(1):130–137, January 2020. ISSN 1556-1380. doi: 10.1016/j.jtho.2019.09.081.

[55] Kelly Dufraing, Gert De Hertogh, Véronique Tack, Cleo Keppens, Elisabeth M. C. Dequeker, and J. Han van Krieken. External Quality Assessment Identifies Training Needs to Determine the Neoplastic Cell Content for Biomarker Testing. *The Journal of molecular diagnostics: JMD*, 20(4):455–464, July 2018. ISSN 1943-7811. doi: 10.1016/j.jmoldx.2018.03.003.

[56] William D. Travis, Megan Eisele, Katherine K. Nishimura, Rania G. Aly, Pietro Bertoglio, Teh-Ying Chou, Frank C. Detterbeck, Jessica Donnington, Wentao Fang, Philippe Joubert, Kemp Kernstine, Young Tae Kim, Yolande Lievens, Hui Liu, Gustavo Lyons, Mari Mino-Kenudson, Andrew G. Nicholson, Mauro Papotti, Ramon Rami-Porta, Valerie Rusch, Shuji Sakai, Paula Ugalde, Paul Van Schil, Chi-Fu Jeffrey Yang, Vanessa J. Cilento, Masaya Yotsukura, Hisao Asamura, and Members of the International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee, Members of the Advisory Boards, and Participating Institutions of

the Lung Cancer Domain. The International Association for the Study of Lung Cancer (IASLC) Staging Project for Lung Cancer: Recommendation to Introduce Spread Through Air Spaces as a Histologic Descriptor in the Ninth Edition of the TNM Classification of Lung Cancer. Analysis of 4061 Pathologic Stage I NSCLC. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 19(7):1028–1051, July 2024. ISSN 1556-1380. doi: 10.1016/j.jtho.2024.03.015.

[57] Kristin A Higgins, Junzo P Chino, Neal Ready, Thomas A D'Amico, Mark F Berry, Thomas Sporn, Jessamy Boyd, and Chris R Kelsey. Lymphovascular Invasion in Non–Small-Cell Lung Cancer: Implications for Staging and Adjuvant Therapy. *Journal of Thoracic Oncology*, 7(7): 1141–1147, July 2012. ISSN 1556-0864. doi: 10.1097/JTO.0b013e3182519a42. URL https://www.sciencedirect.com/science/article/pii/S1556086415332913.

[58] Romain Kessler, Bernard Gasser, Gilbert Massard, Norbert Roeslin, Pierre Meyer, Jean-Marie Wihlm, and Georges Morand. Blood vessel invasion is a major prognostic factor in resected non-small cell lung cancer. *The Annals of Thoracic Surgery*, 62(5):1489–1493, November 1996. ISSN 00034975. doi: 10.1016/0003-4975(96)00540-1. URL https://linkinghub.elsevier.com/retrieve/pii/0003497596005401.

[59] William D. Travis, Sanja Dacic, Ignacio Wistuba, Lynette Sholl, Prasad Adusumilli, Lukas Bubendorf, Paul Bunn, Tina Cascone, Jamie Chaft, Gang Chen, Teh-Ying Chou, Wendy Cooper, Jeremy J. Erasmus, Carlos Gil Ferreira, Jin-Mo Goo, John Heymach, Fred R. Hirsch, Hidehito Horinouchi, Keith Kerr, Mark Kris, Deepali Jain, Young T. Kim, Fernando Lopez-Rios, Shun Lu, Tetsuya Mitsudomi, Andre Moreira, Noriko Motoi, Andrew G. Nicholson, Ricardo Oliveira, Mauro Papotti, Ugo Pastorino, Luis Paz-Ares, Giuseppe Pelosi, Claudia Poleri, Mariano Provencio, Anja C. Roden, Giorgio Scagliotti, Stephen G. Swisher, Erik Thunnissen, Ming S. Tsao, Johan Vansteenkiste, Walter Weder, and Yasushi Yatabe. IASLC Multidisciplinary Recommendations for Pathologic Assessment of Lung Cancer Resection Specimens After Neoadjuvant Therapy. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 15(5):709–740, May 2020. ISSN 1556-1380. doi: 10.1016/j.jtho.2020.01.005.

[60] Marius Ilie, Véronique Hofman, Cécile Ortholan, Christelle Bonnetaud, Céline Coëlle, Jérôme Mouroux, and Paul Hofman. Predictive clinical outcome of the intratumoral CD66b-positive neutrophil-to-CD8-positive T-cell ratio in patients with resectable nonsmall cell lung cancer. *Cancer*, 118(6):1726–1737, 2012. ISSN 1097-0142. doi: 10.1002/cncr.26456. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.26456.

[61] Masayuki Shirasawa, Tatsuya Yoshida, Yukiko Shimoda, Daisuke Takayanagi, Kouya Shiraishi, Takashi Kubo, Sachiyo Mitani, Yuji Matsumoto, Ken Masuda, Yuki Shinno, Yusuke Okuma, Yasushi Goto, Hidehito Horinouchi, Hitoshi Ichikawa, Takashi Kohno, Noboru Yamamoto, Shingo Matsumoto, Koichi Goto, Shun-ichi Watanabe, Yuichiro Ohe, and Noriko Motoi. Differential Immune-Related Microenvironment Determines Programmed Cell Death Protein-1/Programmed Death-Ligand 1 Blockade Efficacy in Patients With Advanced NSCLC. *Journal of Thoracic Oncology*, 16(12):2078–2090, December 2021. ISSN 1556-0864, 1556-1380. doi: 10.1016/j.jtho.2021.07.027. URL https://www.jto.org/article/S1556-0864(21)02373-X/fulltext.

[62] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language

models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2. Publisher: Nature Publishing Group.

[63] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9):100802, September 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100802. Place: New York, N.Y.

[64] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative Radiology Report Evaluation and Error Notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

[65] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, March 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL https://www.nature.com/articles/s41591-024-03423-7. Publisher: Nature Publishing Group.

[66] Carina Kludt, Yuan Wang, Waleed Ahmad, Andrey Bychkov, Junya Fukuoka, Nadine Gaisa, Mark Kühnel, Danny Jonigk, Alexey Pryalukhin, Fabian Mairinger, Franziska Klein, Anne Maria Schultheis, Alexander Seper, Wolfgang Hulla, Johannes Brägelmann, Sebastian Michels, Sebastian Klein, Alexander Quaas, Reinhard Büttner, and Yuri Tolkach. Next-generation lung cancer pathology: Development and validation of diagnostic and prognostic algorithms. *Cell Reports. Medicine*, 5(9):101697, September 2024. ISSN 2666-3791. doi: 10.1016/j.xcrm.2024.101697.

[67] Joey Spronck, Leander van Eekelen, Dominique van Midden, Joep Bogaerts, Leslie Tessier, Valerie Dechering, Muradije Demirel-Andishmand, Gabriel Silva de Souza, Roland Nemeth, Enrico Munari, Giuseppe Bogina, Ilaria Girolami, Albino Eccher, Balazs Acs, Ceren Boyaci, Natalie Klubickova, Monika Looijen-Salamon, Shoko Vos, and Francesco Ciompi. A tissue and cell-level annotated H&E and PD-L1 histopathology image dataset in non-small cell lung cancer, July 2025. URL http://arxiv.org/abs/2507.16855. arXiv:2507.16855 [q-bio].

[68] Benjamin Adjadj, Pierre-Antoine Bannier, Guillaume Horent, Sebastien Mandela, Aurore Lyon, Kathryn Schutte, Ulysse Marteau, Valentin Gaury, Laura Dumont, Thomas Mathieu, MOSAIC consortium, Reda Belbahri, Benoît Schmauch, Eric Durand, Katharina Von Loga, and Lucie Gillet. Towards Comprehensive Cellular Characterisation of H&E slides, September 2025. URL http://arxiv.org/abs/2508.09926. arXiv:2508.09926 [cs.CV].

[69] S. Wu, J. Shang, Z. Li, H. Liu, X. Xu, Z. Zhang, Y. Wang, M. Zhao, M. Yue, J. He, J. Miao, Y. Sang, J. Yan, W. Pang, Q. Shao, Y. Zhang, M. Zhao, X. Liu, P. Wang, C. Cai, B. Liu, X. Wang, and Y. Liu. Interobserver consistency and diagnostic challenges in HER2-ultralow breast cancer: a multicenter

study. *ESMO Open*, 10(2), February 2025. ISSN 2059-7029. doi: 10.1016/j.esmoop.2024.104127. URL https://www.esmoopen.com/article/S2059-7029(24)01898-2/fulltext.

[70] Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, Daniel Shao Weng Tan, Weiwei Zhai, Anders Jacobsen Skanderup, Wing-Kin Sung, and Hwee Kuan Lee. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. *Patterns*, 3(2), February 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100399. URL https://www.cell.com/patterns/abstract/S2666-3899(21)00266-X. Publisher: Elsevier.

[71] Omar S. M. El Nahhas, Chiara M. L. Loeffler, Zunamys I. Carrero, Marko van Treeck, Fiona R. Kolbinger, Katherine J. Hewitt, Hannah S. Muti, Mara Graziani, Qinghe Zeng, Julien Calderaro, Nadina Ortiz-Brüchle, Tanwei Yuan, Michael Hoffmeister, Hermann Brenner, Alexander Brobeil, Jorge S. Reis-Filho, and Jakob Nikolas Kather. Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. *Nature Communications*, 15(1):1253, February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45589-1. URL https://www.nature.com/articles/s41467-024-45589-1. Publisher: Nature Publishing Group.

[72] Sushant Patkar, Alex Chen, Alina Basnet, Amber Bixby, Rahul Rajendran, Rachel Chernet, Susan Faso, Prashanth Ashok Kumar, Devashish Desai, Ola El-Zammar, Christopher Curtiss, Saverio J. Carello, Michel R. Nasr, Peter Choyke, Stephanie Harmon, Baris Turkbey, and Tamara Jamaspishvili. Predicting the tumor microenvironment composition and immunotherapy response in non-small cell lung cancer from digital histopathology images. *npj Precision Oncology*, 8 (1):280, December 2024. ISSN 2397-768X. doi: 10.1038/s41698-024-00765-w. URL https://www.nature.com/articles/s41698-024-00765-w. Publisher: Nature Publishing Group.

[73] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative AI copilot for human pathology. *Nature*, 634(8033):466–473, October 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07618-3. URL https://www.nature.com/articles/s41586-024-07618-3.

[74] Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H. Bernhard, Ran A. Godrich, Juan A. Retamero, Jinru Shia, Mithat Gonen, Martin R. Weiser, David S. Klimstra, Razik Yousfi, Nicolò Fusi, Thomas J. Fuchs, Kristen Severson, and Siqi Liu. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, pages 1–11, July 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04521-4. URL https://www.nature.com/articles/s41591-026-04521-4. Publisher: Nature Publishing Group.

[75] Rundong Wang, Wei Ba, Ying Zhou, Yingtai Li, Bowen Liu, Baizhi Wang, Yuhao Wang, Zhidong Yang, Kun Zhang, Rui Yan, and S. Kevin Zhou. QCAgent: An agentic framework for quality-controllable pathology report generation from whole slide image, March 2026. URL http://arxiv.org/abs/2603.01647. arXiv:2603.01647 [cs.CV].

[76] Zhe Xu, Zhengyu Zhang, Zhiyuan Cai, Jiahao Xu, Yijie Lin, Ziyi Liu, Junlin Hou, Hongyi Wang, Yuxiang Nie, Yihui Wang, Jiabo Ma, Ling Liang, Yingxue Xu, Zhengrui Guo, Guanghao Wu, Danyi Li, Ziqi Zhou, Donglin Tan, Zhijian Cen, Ying Tan, Xiaolin Liu, Qi Xie, Xiaoying Tang, Xi Peng, Cheng Deng, Lijuan Qu, Ronald Cheong Kin Chan, Li Liang, and Hao Chen. A Multimodal Agentic Pathology Co-pilot via Evidence Grounded Reasoning, August 2026. URL http://arxiv.org/abs/2606.08093. arXiv:2606.08093 [cs.AI].

[77] Giulio Rossi, Alessandro Marchioni, Marina Milani, Rosa Scotti, Moira Foroni, AnnaMaria Cesinaro, Lucio Longo, Mario Migaldi, and Alberto Cavazza. TTF-1, cytokeratin 7, 34betaE12, and CD56/NCAM immunostaining in the subclassification of large cell carcinomas of the lung. *American Journal of Clinical Pathology*, 122(6):884–893, December 2004. ISSN 0002-9173.

[78] Mi Jin Kim, Hyeong Chan Shin, Kyeong Cheol Shin, and Jae Y. Ro. Best immunohistochemical panel in distinguishing adenocarcinoma from squamous cell carcinoma of lung: tissue microarray assay in resected lung cancer specimens. *Annals of Diagnostic Pathology*, 17(1):85–90, February 2013. ISSN 1092-9134. doi: 10.1016/j.anndiagpath.2012.07.006. URL https://www.sciencedirect.com/science/article/pii/S1092913412001037.

[79] Yasushi Yatabe, Sanja Dacic, Alain C. Borczuk, Arne Warth, Prudence A. Russell, Sylvie Lantuejoul, Mary Beth Beasley, Erik Thunnissen, Giuseppe Pelosi, Natasha Rekhtman, Lukas Bubendorf, Mari Mino-Kenudson, Akihiko Yoshida, Kim R. Geisinger, Masayuki Noguchi, Lucian R. Chirieac, Johan Bolting, Jin-Haeng Chung, Teh-Ying Chou, Gang Chen, Claudia Poleri, Fernando Lopez-Rios, Mauro Papotti, Lynette M. Sholl, Anja C. Roden, William D. Travis, Fred R. Hirsch, Keith M. Kerr, Ming-Sound Tsao, Andrew G. Nicholson, Ignacio Wistuba, and Andre L. Moreira. Best Practices Recommendations for Diagnostic Immunohistochemistry in Lung Cancer. *Journal of Thoracic Oncology*, 14(3):377–407, March 2019. ISSN 1556-0864. doi: 10.1016/j.jtho.2018.12.005. URL https://www.sciencedirect.com/science/article/pii/S1556086418335147.

[80] Philipp Anders, Marvin Sextro, Katja Lingelbach, Kai Standvoss, Suhas Pandhe, Sandip Ghosh, Cornelius Böhm, Stephan Tietz, Rosemarie Krupar, Lars Tharun, Marie-Lisa Eich, Julika Ribbat-Idel, Evelyn Ramberger, Xizi Liang, Verena Aumiller, Sabine Merkelbach-Bruse, Alexander Quaas, Nikolaj Frost, Georg Schlachtenberger, Matthias Heldwein, Ulrich Keilholz, Khosro Hekmat, Jens-Carsten Rückert, Reinhard Büttner, Christian Grohe, David Horst, Maximilian Alber, Lukas Ruff, Frederick Klauschen, Gabriel Dernbach, Philipp Seegerer, and Simon Schallenberg. ADC Target Profiling in NSCLC: Generalizable AI Separates TROP-2 and cMET Phenotypes. *Clinical Cancer Research*, pages OF1–OF22, May 2026. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-25-4513. URL https://doi.org/10.1158/1078-0432.CCR-25-4513.

[81] Tony S K Mok, Yi-Long Wu, Iveta Kudaba, Dariusz M Kowalski, Byoung Chul Cho, Hande Z Turna, Gilberto Castro, Vichien Srimuninnimit, Konstantin K Laktionov, Igor Bondarenko, Kaoru Kubota, Gregory M Lubiniecki, Jin Zhang, Debra Kush, Gilberto Lopes, Grigory Adamchuk, Myung-Ju Ahn, Aurelia Alexandru, Ozden Altundag, Anna Alyasova, Orest Andrusenko, Keisuke Aoe, Antonio Araujo, Osvaldo Aren, Oscar Arrieta Rodriguez, Touch Ativitavas, Oscar Avendano, Fernando Barata, Carlos Henrique Barrios, Carlos Beato, Per Bergstrom, Daniel Betticher, Larisa Bolotina, Igor Bondarenko, Michiel Botha, Sayeuri Buddu, Christian Caglevic, Andres Cardona, Gilberto Castro, Hugo Castro, Filiz Cay Senler, Carlos Alexandre Sydow Cerny, Alvydas Cesas, Gee-Chen Chan, Jianhua Chang, Gongyan Chen, Xi Chen, Susanna Cheng, Ying Cheng, Nelly Cherciu, Chao-Hua Chiu, Byoung Chul Cho, Saulius Cicenas, Daniel Ciurescu, Graham Cohen, Marcos Andre Costa, Pongwut Danchaivijitr, Flavia De Angelis, Sergio Jobim de Azevedo, Mircea Dediu, Tsvetan Deliverski, Pedro Rafael Martins De Marchi, Flor de The Bustamante Valles, Zhenyu Ding, Boyan Doganov, Lydia Dreosti, Ricardo Duarte, Regina Edusma-Dy, Sergey Emelyanov, Mustafa Erman, Yun Fan, Luis Fein, Jifeng Feng, David Fenton, Gustavo Fernandes, Carlos Ferreira, Fabio Andre Franke, Helano Freitas, Yasuhito Fujisaka, Hector Galindo, Christina Galvez, Doina Ganea, Nuno Gil, Gustavo Girotto, Erdem Goker, Tuncay Goksel, Gonzalo Gomez Aubin, Luis Gomez Wolff, Hakan Griph, Mahmut Gumus, Jacqueline Hall, Gregory Hart, Libor Havel, Jianxing He, Yong He, Carlos Hernandez Hernandez, Venceslau Hespanhol, Tomonori Hirashima, Chung Man James Ho, Atsushi Horiike, Yukio Hosomi, Katsuyuki Hotta, Mei Hou, Soon Hin How, Te-Chun Hsia, Yi Hu, Masao Ichiki, Fumio Imamura, Oleksandr Ivashchuk, Yasuo Iwamoto, Jana Jaal, Jacek

Jassem, Christa Jordaan, Rosalyn Anne Juergens, Diego Kaen, Ewa Kalinka-Warzocha, Nina Karaseva, Boguslawa Karaszewska, Andrzej Kazarnowicz, Kazuo Kasahara, Nobuyuki Katakami, Terufumi Kato, Tomoya Kawaguchi, Joo Hang Kim, Kazuma Kishi, Vitezslav Kolek, Marchela Koleva, Petr Kolman, Leona Koubkova, Ruben Kowalyszyn, Dariusz Kowalski, Krassimir Koynov, Doran Ksienski, Kaoru Kubota, Iveta Kudaba, Takayasu Kurata, Gerli Kuusk, Lyudmila Kuzina, Ibolya Laczo, Guia Elena Imelda Ladrera, Konstantin Laktionov, Gregory Landers, Sergey Lazarev, Guillermo Lerzo, Krzysztof Lesniewski Kmak, Wei Li, Chong Kin Liam, Igor Lifirenko, Oleg Lipatov, Xiaoqing Liu, Zhe Liu, Sing Hung Lo, Valeria Lopes, Karla Lopez, Shun Lu, Gaston Martinengo, Luis Mas, Marina Matrosova, Rumyana Micheva, Zhasmina Milanova, Lucian Miron, Tony Mok, Matias Molina, Shuji Murakami, Yasuharu Nakahara, Tien Quang Nguyen, Takashi Nishimura, Adrian Ochsenbein, Tatsuo Ohira, Ronny Ohman, Choo Khoon Ong, Gyula Ostoros, Xuenong Ouyang, Elena Ovchinnikova, Ozgur Ozyilkan, Lubos Petruzelka, Xuan Dung Pham, Pablo Picon, Bela Piko, Artem Poltoratsky, Olga Ponomarova, Patrice Popelkova, Gunta Purkalne, Shukui Qin, Rodryg Ramlau, Bernardo Rappaport, Felipe Rey, Eduardo Richardet, Jaromir Roubec, Paul Ruff, Andrii Rusyn, Hideo Saka, Jorge Salas, Mario Sandoval, Lucas Santos, Toshiyuki Sawa, Kasan Seetalarom, Mesut Seker, Nobuhiko Seki, Freddy Seolwane, Lucinda Shepherd, Sergii Shevnya, Andrea Kazumi Shimada, Yaroslav Shparyk, Ivan Sinielnikov, Daniela Sirbu, Oren Smaletz, Joao Paulo Holanda Soares, Aumkhae Sookprasert, Giovanna Speranza, Vichien Srimuninnimit, Virote Sriuranpong, Zinaida Stara, Wu-Chou Su, Shunichi Sugawara, Waldemar Szpak, Kazuhisa Takahashi, Nagio Takigawa, Hiroshi Tanaka, Jerry Tan Chun Bing, Qiyou Tang, Pavel Taranov, Hermes Tejada, Lye Mun Tho, Yoshitaro Torii, Dmytro Trukhyn, Maria Turdean, Hande Turna, Grygoriy Ursol, Jaroslav Vanasek, Mirta Varela, Marcela Vallejo, Luis Vera, Ana-Paula Victorino, Tomas Vlasek, Ihor Vynnychenko, Buhai Wang, Jie Wang, Kai Wang, Yilong Wu, Kazuhiko Yamada, Chih-Hsin Yang, Takuma Yokoyama, Toshihide Yokoyama, Hiroshige Yoshioka, Fulden Yumuk, Angela Zambrano, Juan Jose Zarba, Oleg Zarubenkov, Marius Zemaitis, Li Zhang, Li Zhang, Xin Zhang, Jun Zhao, Caicun Zhou, Jianying Zhou, Qing Zhou, and Alfred Zippelius. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. *The Lancet*, 393(10183):1819–1830, May 2019. ISSN 0140-6736. doi: 10.1016/S0140-6736(18)32409-7. URL https://www.sciencedirect.com/science/article/pii/S0140673618324097.

[82] Bruno da Silveira Corrêa, Fernanda De-Paris, Guilherme Danielski Viola, Tiago Finger Andreis, Clévia Rosset, Fernanda Sales Luiz Vianna, Luis Fernando da Rosa Rivero, Francine Hehn de Oliveira, Patricia Ashton-Prolla, and Gabriel de Souza Macedo. Challenges to the effectiveness of next-generation sequencing in formalin-fixed paraffin-embedded tumor samples for non-small cell lung cancer. *Annals of Diagnostic Pathology*, 69:152249, April 2024. ISSN 1092-9134. doi: 10.1016/j.anndiagpath.2023.152249. URL https://www.sciencedirect.com/science/article/pii/S1092913423001478.

[83] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology*, 125(8):1264–1272, August 2018. ISSN 1549-4713. doi: 10.1016/j.ophtha.2018.01.034.

[84] Kimberly H. Allison, Lisa M. Reisch, Patricia A. Carney, Donald L. Weaver, Stuart J. Schnitt, Frances P. O'Malley, Berta M. Geller, and Joann G. Elmore. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology*, 65(2):240–251, August 2014. ISSN 1365-2559. doi: 10.1111/his.12387.

[85] Benoît Lhermitte, Caroline Egele, Noëlle Weingertner, Damien Ambrosetti, Bérengère Dadone, Valérie Kubiniek, Fanny Burel-Vandenbos, John Coyne, Jean-François Michiels, Marie-Pierre

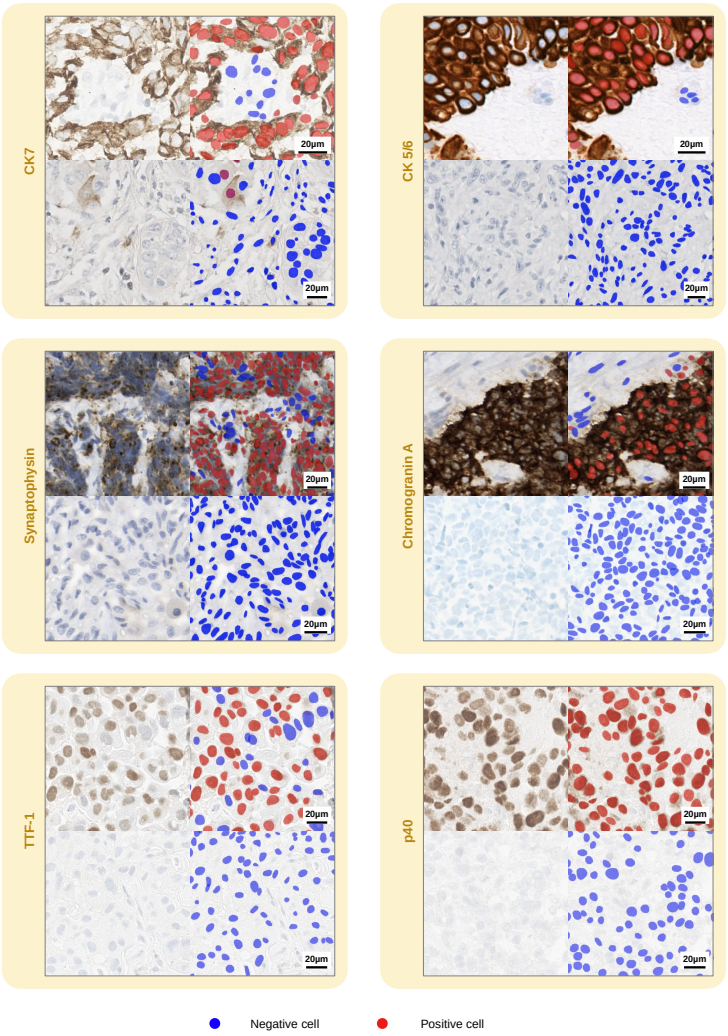
Chenard, Etienne Rouleau, Jean-Christophe Sabourin, and Jean-Pierre Bellocq. Adequately defining tumor cell proportion in tissue samples for molecular testing improves interobserver reproducibility of its assessment. *Virchows Archiv: An International Journal of Pathology*, 470(1): 21–27, January 2017. ISSN 1432-2307. doi: 10.1007/s00428-016-2042-6.

[86] Taro Sakamoto, Tomoi Furukawa, Kris Lami, Hoa Hoang Ngoc Pham, Wataru Uegami, Kishio Kuroda, Masataka Kawai, Hidenori Sakanashi, Lee Alex Donald Cooper, Andrey Bychkov, and Junya Fukuoka. A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. *Translational Lung Cancer Research*, 9(5):2255–2276, October 2020. ISSN 2218-6751. doi: 10.21037/tlcr-20-591.

[87] Tomoharu Kiyuna, Eric Cosatto, Kanako C. Hatanaka, Tomoyuki Yokose, Koji Tsuta, Noriko Motoi, Keishi Makita, Ai Shimizu, Toshiya Shinohara, Akira Suzuki, Emi Takakuwa, Yasunari Takakuwa, Takahiro Tsuji, Mitsuhiro Tsujiwaki, Mitsuru Yanai, Sayaka Yuzawa, Maki Ogura, and Yutaka Hatanaka. Evaluating Cellularity Estimation Methods: Comparing AI Counting with Pathologists' Visual Estimates. *Diagnostics*, 14(11):1115, January 2024. ISSN 2075-4418. doi: 10.3390/diagnostics14111115. URL https://www.mdpi.com/2075-4418/14/11/1115.

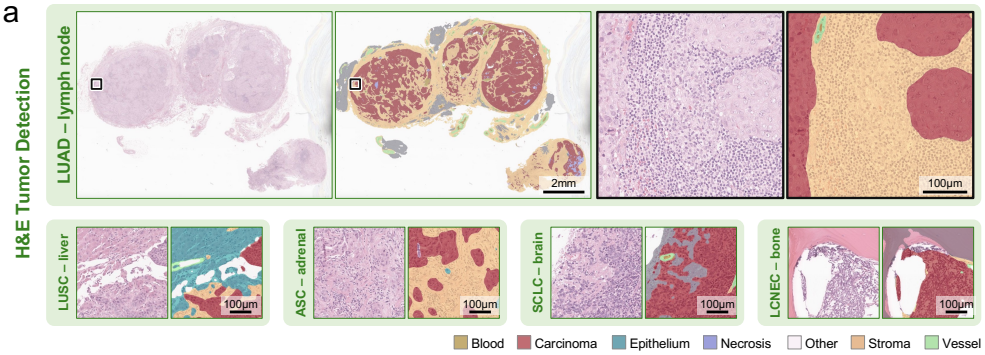
[88] Arkadiusz Gertych, Natalia Zurek, Natalia Piaseczna, Kamil Szkaradnik, Yujie Cui, Yi Zhang, Karolina Nurzynska, Bartłomiej Pyciński, Piotr Paul, Artur Bartczak, Ewa Chmielik, and Ann E. Walts. Tumor Cellularity Assessment Using Artificial Intelligence Trained on Immunohistochemistry-Restained Slides Improves Selection of Lung Adenocarcinoma Samples for Molecular Testing. *The American Journal of Pathology*, 195(5):907–922, May 2025. ISSN 1525-2191. doi: 10.1016/j.ajpath.2025.01.009.

# A Supplementary Tables And Supplementary Figures




Supplementary Figure 1: **Immunohistochemical Tumor Subtyping.** Representative examples of AI model expression scoring for cytoplasmic markers CK7, CK5/6, synaptophysin, and chromogranin A, and nuclear markers TTF-1 and p40. For each marker, a positive case (upper two panels) and a negative case (lower two panels) are shown, each comprising the AI model cell classification overlay (right panels) and the corresponding unmodified IHC image (left panels). For cytoplasmic markers, positivity was determined based on LUCAID module 3 cytoplasmic biomarker expression scoring; for nuclear markers TTF-1 and p40, positivity was determined by nuclear DAB intensity thresholding. Cases with more than 10% positive tumor cells were classified as marker-positive.
*Abbreviations:* CK5/6: cytokeratin 5/6; CK7: cytokeratin 7; DAS: 3,3'-Diaminobenzidin, IHC, immunohistochemical; TTF-1: thyroid transcription factor 1.

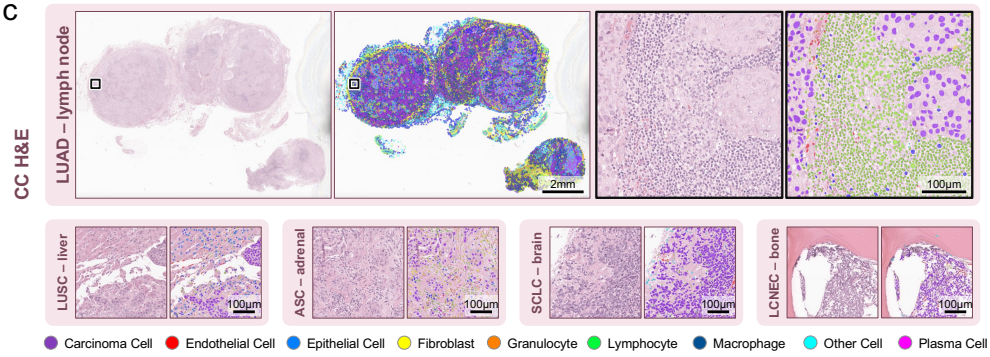
a




b

| Prediction Class | Carcinoma | Epithelium | Stroma | Necrosis | Blood | Vessel | Other | Average |
|---|---|---|---|---|---|---|---|---|
| **F1 Overall** | 0.99 | 0.98 | 0.93 | 0.97 | 0.96 | 0.78 | 0.98 | **0.94** |
| **F1 LUAD** | 0.99 | 0.99 | 0.91 | 0.92 | 0.89 | 0.83 | 0.91 | **0.92** |
| **F1 LUSC** | 0.99 | 1.00 | 0.87 | 0.97 | 0.96 | 0.61 | 0.91 | **0.90** |
| **F1 ASC** | 0.99 | 0.93 | 0.97 | 0.60 | 0.97 | 0.96 | 1.00 | **0.92** |
| **F1 SCLC** | 0.97 | 0.98 | 0.67 | 0.93 | 0.94 | 0.78 | 0.99 | **0.89** |
| **F1 LCNEC** | 0.97 | 0.82 | 0.89 | 0.98 | 0.98 | 0.61 | 0.97 | **0.89** |

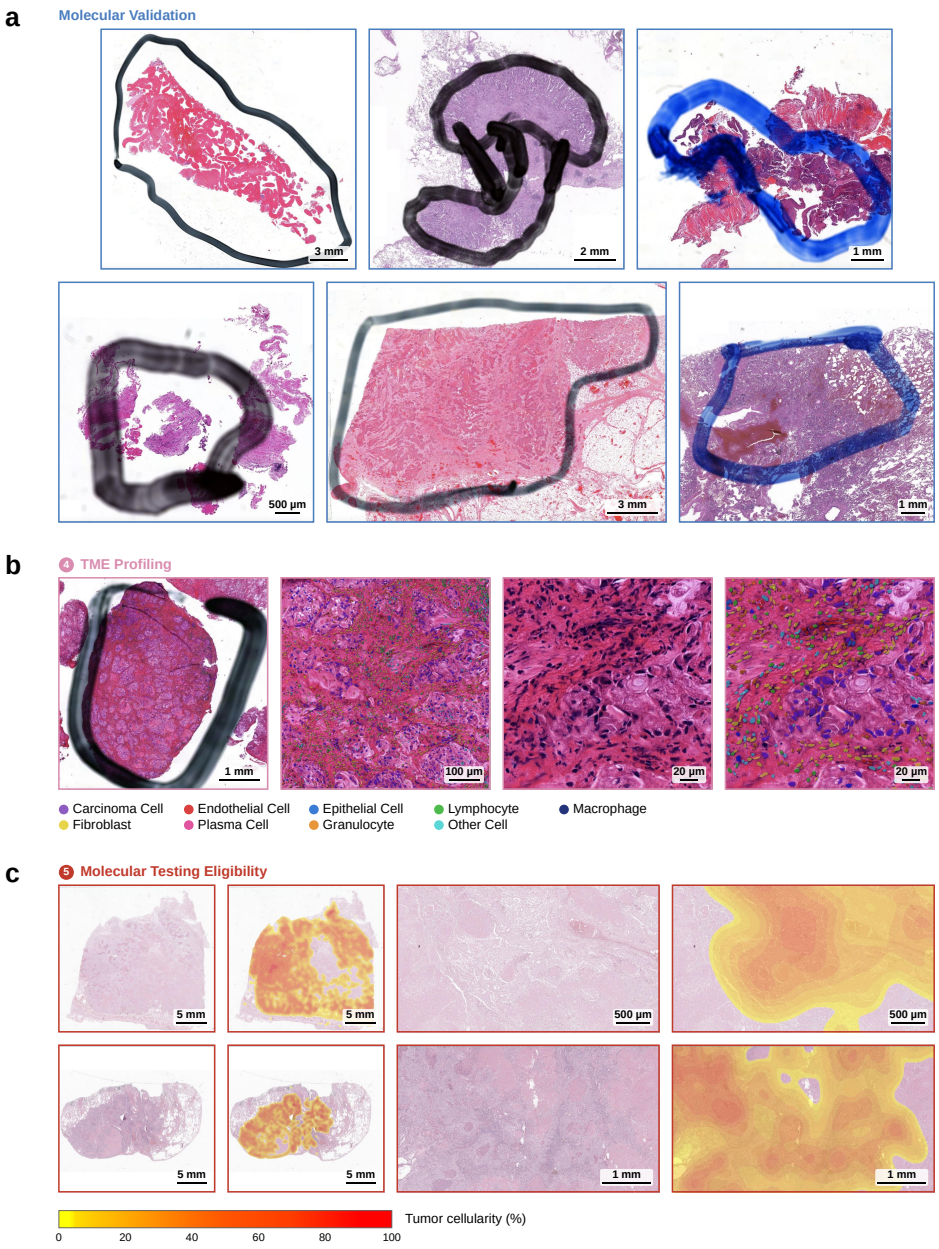
c



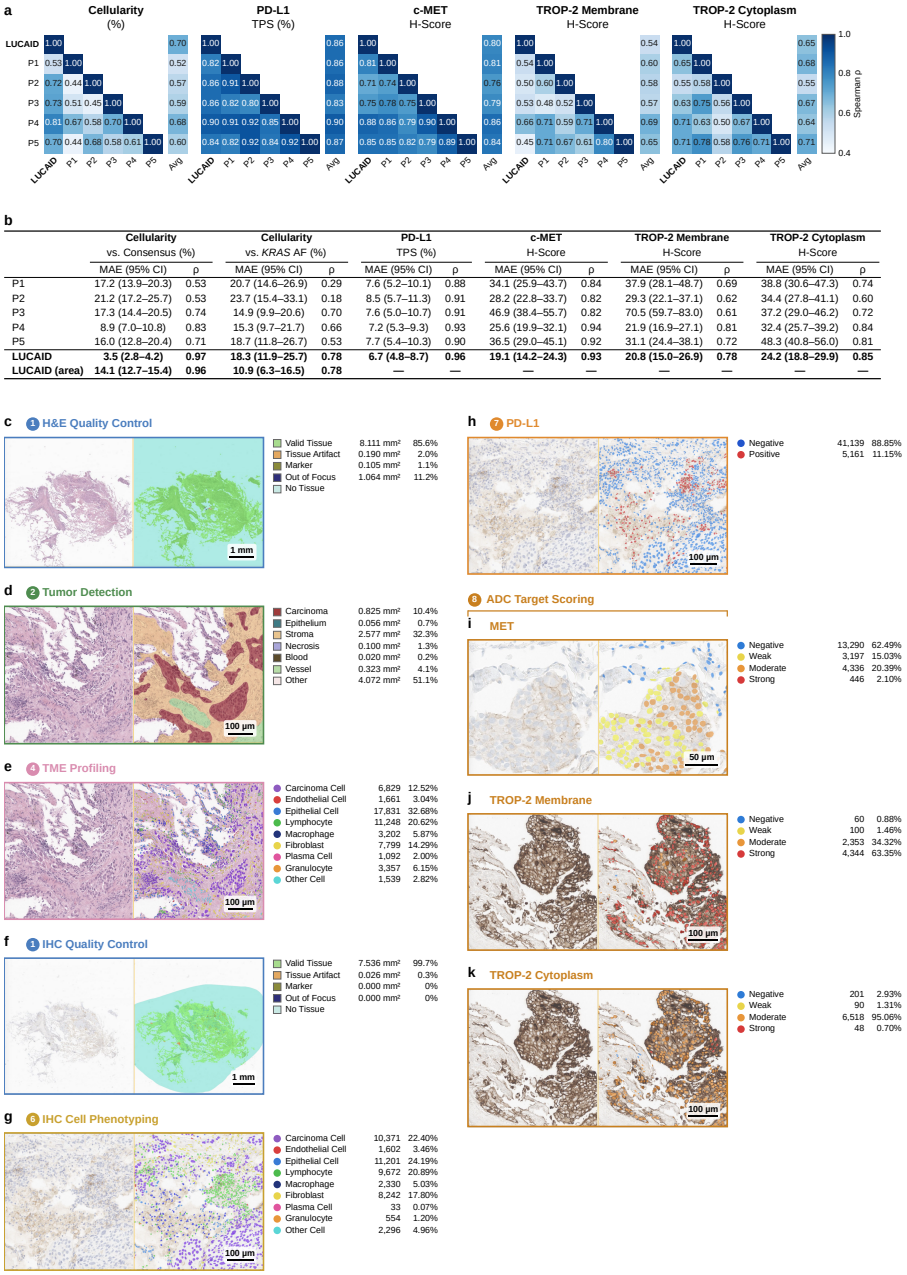

d

| Prediction Class | CaC | EC | Fib | Lym | PC | GC | M | EnC | Other Cell | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **F1 Overall** | 0.99 | 0.98 | 0.92 | 0.91 | 0.88 | 0.90 | 0.75 | 0.92 | 0.93 | **0.91** |
| **F1 LUAD** | 0.98 | 0.97 | 0.89 | 0.91 | 0.80 | 0.83 | 0.81 | 0.86 | 0.86 | **0.88** |
| **F1 LUSC** | 0.98 | 1.00 | 0.97 | 0.93 | 0.89 | 0.91 | 0.78 | 0.99 | 0.93 | **0.93** |
| **F1 ASC** | 0.98 | 0.92 | 0.78 | 0.90 | 0.84 | 0.99 | 0.72 | 0.90 | 0.79 | **0.87** |
| **F1 SCLC** | 1.00 | 1.00 | 0.89 | 0.80 | 0.56 | 0.91 | 0.60 | 0.92 | 0.91 | **0.84** |
| **F1 LCNEC** | 0.99 | 0.96 | 0.94 | 0.92 | 0.87 | 0.90 | 0.67 | 0.93 | 0.93 | **0.90** |

Supplementary Figure 2: **LUCAID H&E Tumor Detection and Cell Phenotyping in Metastatic Lung Cancer Tumors. a**, H&E tumor-detection examples in metastases from five major lung carcinoma subtypes, with histopathological images shown on the left and corresponding model-derived tissue-segmentation overlays on the right. For LUAD, a lymph node metastasis is shown as an overview image with a higher-magnification view; the lower row shows metastases of LUSC (liver), ASC (adrenal gland), SCLC (brain) and LCNEC (bone) from left to right. The segmentation model distinguishes seven tissue compartments: carcinoma, epithelium, stroma, necrosis, blood, vessel and other. **b**, F1 scores for each tissue compartment overall and stratified by histological subtype. **c**, H&E cell-phenotyping examples in the same metastatic specimens, with histopathological images shown on the left and corresponding model-derived cell-classification overlays on the right. The model distinguishes nine cell classes: carcinoma cells, endothelial cells, epithelial cells, fibroblasts, granulocytes, lymphocytes, macrophages, plasma cells and other cells. **d**, F1 scores for each cell class overall and stratified by histological subtype.
Abbreviations: ASC, adenosquamous carcinoma; CaC, carcinoma cell; CC, cell classification; EC, endothelial cell; EnC, epithelial cell; Fib, fibroblast; GC, granulocyte; H&E, hematoxylin and eosin; LCNEC, large cell neuroendocrine carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; Lym, lymphocyte; M, macrophage; PC, plasma cell; SCLC, small cell lung cancer.



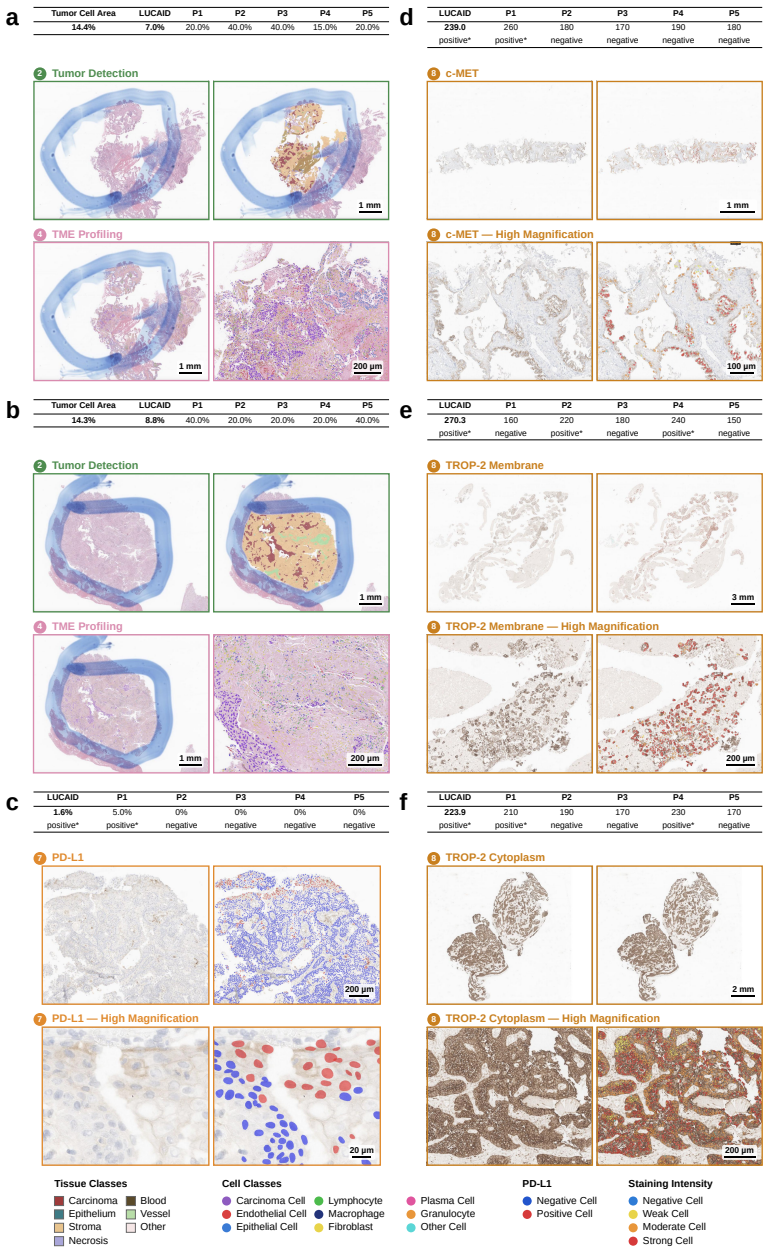

Supplementary Figure 3: **Molecular Validation of LUCAID Cell Phenotyping module in Lung Cancer Patients.** **a**, Pathologist pen-marked H&E whole slide images in six representative cases, each shown as a whole-slide overview with the pen-marked region designating the area selected for molecular analysis. **b**, TME profiling (module 4) in a representative lung adenocarcinoma. Whole-slide overview with the pen-marked tumor area (left), model overlay at intermediate magnification (second from left) and high magnification shown as original image and corresponding model overlay (two right panels). The model distinguishes nine cell classes: carcinoma cells, endothelial cells, epithelial cells, fibroblasts, granulocytes, lymphocytes, macrophages, plasma cells and other cells. **c**, Molecular testing eligibility (module 5) in two representative cases (rows). Each case is shown as original image and corresponding tumor cellularity heatmap at whole-slide overview (left pair) and high magnification (right pair). Color encodes predicted tumor cellularity (%). *Abbreviations:* H&E, haematoxylin and eosin; TME, tumor microenvironment.




**b**

| | Cellularity vs. Consensus (%) | | Cellularity vs. *KRAS* AF (%) | | PD-L1 TPS (%) | | c-MET H-Score | | TROP-2 Membrane H-Score | | TROP-2 Cytoplasm H-Score | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE (95% CI) | ρ | MAE (95% CI) | ρ | MAE (95% CI) | ρ | MAE (95% CI) | ρ | MAE (95% CI) | ρ | MAE (95% CI) | ρ |
| P1 | 17.2 (13.9–20.3) | 0.53 | 20.7 (14.6–26.9) | 0.29 | 7.6 (5.2–10.1) | 0.88 | 34.1 (25.9–43.7) | 0.84 | 37.9 (28.1–48.7) | 0.69 | 38.8 (30.6–47.3) | 0.74 |
| P2 | 21.2 (17.2–25.7) | 0.53 | 23.7 (15.4–33.1) | 0.18 | 8.5 (5.7–11.3) | 0.91 | 28.2 (22.8–33.7) | 0.82 | 29.3 (22.1–37.1) | 0.62 | 34.4 (27.8–41.1) | 0.60 |
| P3 | 17.3 (14.4–20.5) | 0.74 | 14.9 (9.9–20.6) | 0.70 | 7.6 (5.0–10.7) | 0.91 | 46.9 (38.4–55.7) | 0.82 | 70.5 (59.7–83.0) | 0.61 | 37.2 (29.0–46.2) | 0.72 |
| P4 | 8.9 (7.0–10.8) | 0.83 | 15.3 (9.7–21.7) | 0.66 | 7.2 (5.3–9.3) | 0.93 | 25.6 (19.9–32.1) | 0.94 | 21.9 (16.9–27.1) | 0.81 | 32.4 (25.7–39.2) | 0.84 |
| P5 | 16.0 (12.8–20.4) | 0.71 | 18.7 (11.8–26.7) | 0.53 | 7.7 (5.4–10.3) | 0.90 | 36.5 (29.0–45.1) | 0.92 | 31.1 (24.4–38.1) | 0.72 | 48.3 (40.8–56.0) | 0.81 |
| **LUCAID** | **3.5 (2.8–4.2)** | **0.97** | **18.3 (11.9–25.7)** | **0.78** | **6.7 (4.8–8.7)** | **0.96** | **19.1 (14.2–24.3)** | **0.93** | **20.8 (15.0–26.9)** | **0.78** | **24.2 (18.8–29.9)** | **0.85** |
| **LUCAID (area)** | **14.1 (12.7–15.4)** | **0.96** | **10.9 (6.3–16.5)** | **0.78** | — | — | — | — | — | — | — | — |

Supplementary Figure 4: **Supplementary Figure 4: Extended prospective clinical validation across raters and pipeline modules. a**, Pairwise Spearman correlation matrices for LUCAID and five pathologists (P1–P5) across tumor cellularity, PD-L1 tumor proportion score (TPS), MET H-score, TROP-2 membranous H-score and TROP-2 cytoplasmic H-score. Color encodes the Spearman correlation coefficient ($\rho$). The rightmost column (Avg) shows each rater's mean pairwise correlation with all other raters. **b**, Mean absolute error (MAE) with 95% confidence intervals and corresponding correlation coefficients for LUCAID and each pathologist across tumor cellularity relative to the expert-panel adjudicated reference standard, tumor cellularity relative to *KRAS* allele frequency, PD-L1 TPS, MET H-score, TROP-2 membranous H-score and TROP-2 cytoplasmic H-score. LUCAID (area) denotes area-based rather than cell-based tumor cellularity estimation. **c**–**k**, Representative LUAD case showing sequential LUCAID module outputs, with histopathological images shown alongside the corresponding model-derived overlays and the associated quantitative readouts. **c**, H&E quality control. **d**, Tumor detection. **e**, TME profiling. **f**, IHC quality control. **g**, IHC cell phenotyping. **h**, PD-L1 scoring. **i**–**k**, ADC target scoring for MET (**i**), TROP-2 membranous (**j**) and TROP-2 cytoplasmic (**k**) expression.
*Abbreviations:* ADC, antibody–drug conjugate; CI, confidence interval; H&E, hematoxylin and eosin; IHC, immunohistochemical; MAE, mean absolute error; MET, hepatocyte growth factor receptor; PD-L1, programmed death-ligand 1; QC, quality control; TME, tumor microenvironment; TPS, tumor proportion score; TROP-2, trophoblast cell-surface antigen 2.




Supplementary Figure 5: **Representative discordant cases across tumor cell content estimation and IHC biomarker scoring.** Tables report LUCAID and pathologist (P1–P5) scores; asterisks denote scores above the clinical positivity threshold. Circled numbers indicate the generating LUCAID module. **a,b**, Tumor cell content. LUCAID estimated tumor cellularity at 7.0% (**a**) and 8.8% (**b**), while all pathologists gave higher estimates (15.0–40.0%). Tumor Detection (module 2, upper) and TME Profiling (module 4, lower) outputs are shown as original image (left) and model overlay (right); high-magnification views show sparse carcinoma cells among predominantly stromal and inflammatory components. **c**, PD-L1 (module 7). LUCAID (TPS 1.6%) and P1 (5.0%) scored above the 1% threshold; P2–P5 scored 0%. **d**, c-MET (module 8). LUCAID (H-score 239.0) and P1 (260) scored above the positivity threshold (H-score $\geq$ 200); P2–P5 scored 170–190. **e**, TROP-2 membranous (module 8). LUCAID (270.3), P2 (220) and P4 (240) scored positive; P1 (160), P3 (180) and P5 (150) negative. **f**, TROP-2 cytoplasmic (module 8). LUCAID (223.9), P1 (210) and P4 (230) scored positive; P2 (190), P3 (170) and P5 (170) negative.
*Abbreviations:* c-MET, hepatocyte growth factor receptor; P1–P5, pathologists 1–5; PD-L1, programmed death-ligand 1; TME, tumor microenvironment; TPS, tumor proportion score; TROP-2, tumor-associated calcium signal transducer 2.

Supplementary Table 1: Clinicopathological Characteristics of the Prospective Clinical Validation Cohort

| **Characteristics** | | **Total n (%)** |
|---|---|---|
| All Patients | | 70 |
| Median Age (years) | | 63 (range 39–84) |
| Gender | Female | 33 (47.1%) |
| | Male | 37 (52.9%) |
| Subtype | Lung adenocarcinoma | 47 (67.1%) |
| | Lung squamous cell carcinoma | 20 (28.6%) |
| | Large cell neuroendocrine carcinoma | 1 (1.4%) |
| | NSCLC NOS[a] | 2 (2.9%) |
| Localization | Primary tumor | 39 (55.7%) |
| | Lymph node metastasis | 14 (20.0%) |
| | Distant metastasis | 17 (24.3%) |
| Specimen type | Biopsy | 48 (68.6%) |
| | Fine needle aspiration | 16 (22.9%) |
| | Resection | 6 (8.6%) |

[a] NSCLC NOS, non-small cell lung carcinoma, not otherwise specified.

Supplementary Table 2: **Supplementary Table 2: Clinicopathological Characteristics of the Discovery Cohort.**

| Parameter | Total $n = 1{,}001$ | AC (LUAD) $n = 581$ | SCC (LUSC) $n = 402$ | ASC $n = 18$ |
|---|---|---|---|---|
| *Sex* | | | | |
| Male | 616 | 308 | 300 | 8 |
| Female | 385 | 273 | 102 | 10 |
| *Age, years* | | | | |
| Median (IQR) | 66 (59–73) | 65 (58–72) | 68 (61–74) | 74 (68–76) |
| *Smoking status* | | | | |
| Never-smoker | 67 | 53 | 12 | 2 |
| Non-smoker | 315 | 184 | 125 | 6 |
| Smoker | 327 | 198 | 123 | 6 |
| Not recorded | 292 | 146 | 142 | 4 |
| *pT stage* | | | | |
| T0/Tx | 3 | 1 | 2 | 0 |
| T1 | 391 | 259 | 127 | 5 |
| T2 | 317 | 188 | 121 | 8 |
| T3 | 180 | 85 | 94 | 1 |
| T4 | 109 | 48 | 57 | 4 |
| Not recorded | 1 | 0 | 1 | 0 |
| *pN stage* | | | | |
| N0 | 633 | 376 | 245 | 12 |
| N1 | 186 | 79 | 105 | 2 |
| N2 | 150 | 105 | 41 | 4 |
| N3 | 11 | 7 | 4 | 0 |
| Not recorded | 21 | 14 | 7 | 0 |
| *pM stage* | | | | |
| M0 | 914 | 512 | 385 | 17 |
| M1a | 14 | 7 | 7 | 0 |
| M1b | 69 | 59 | 9 | 1 |
| M1c | 4 | 3 | 1 | 0 |
| *UICC8 stage* | | | | |
| I | 446 | 274 | 163 | 9 |
| II | 226 | 108 | 115 | 3 |
| III | 242 | 130 | 107 | 5 |
| IV | 87 | 69 | 17 | 1 |
| *Grade* | | | | |
| G1 | 51 | 50 | 1 | 0 |
| G2 | 527 | 309 | 210 | 8 |
| G3 | 303 | 156 | 140 | 7 |
| Not recorded | 120 | 66 | 51 | 3 |
| *Resection status* | | | | |
| R0 | 919 | 539 | 363 | 17 |
| R1 | 81 | 42 | 38 | 1 |
| R2 | 1 | 0 | 1 | 0 |
| *Overall survival* | | | | |
| Deaths | 528 | 272 | 241 | 15 |
| Follow-up, months, median (reverse KM) | 73 | 73 | 75 | 76 |

*Abbreviations:* AC, adenocarcinoma; ASC, adenosquamous carcinoma; IQR, interquartile range; KM, Kaplan–Meier; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; SCC, squamous cell carcinoma; UICC8, Union for International Cancer Control staging system, 8th edition.

Supplementary Table 3: Forest Results of AI-quantified Spatial Features with Feature Definitions

| feature | category | formula | unit | HR | CI_95 | p_value | q_value_BH | split_rule | split_cutoff | n | events |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Carcinoma–plasma cell adjacency | Novel spatial | fraction of neighbour-graph (delaunay) edges linking carcinoma to plasma cells | fraction | 0.75 | 0.63–0.88 | <0.001 | 0.009 | Median | 0.0024 | 1001 | 528 |
| Lymphoid niche (TLS-like) | Novel spatial | fraction of cells in lymphoid-typed leiden communities (lymphocyte + plasma ≥ 40%; tertiary-lymphoid-structure–like aggregate) | fraction | 0.75 | 0.63–0.89 | <0.001 | 0.009 | Gaussian mixture | 0.0531 | 1001 | 528 |
| Stromal TIL % | Established | cell_percentage_lymphocyte_stroma | % | 0.75 | 0.63–0.89 | 0.001 | 0.009 | Median | 18.3725 | 1001 | 528 |
| Plasma cells % | Composition | plasma /total cells | % | 0.77 | 0.65–0.92 | 0.003 | 0.020 | Median | 2.6323 | 1001 | 528 |
| Lympho/granulocyte ratio | Established | dens(lymphocyte)/dens(granulocyte) | ratio | 0.78 | 0.65–0.92 | 0.004 | 0.020 | Median | 3.7804 | 1001 | 528 |
| Macrophage–carcinoma cell distance | Novel spatial | median distance (µm) from each macrophage within the carcinoma compartment to its nearest carcinoma cell (larger = macrophage exclusion) | µm | 0.79 | 0.66–0.93 | 0.006 | 0.029 | Median | 13.8624 | 1001 | 528 |
| Stromal cell density | Established | sum stromal-cell densities | cells/mm$^2$ | 0.79 | 0.67–0.94 | 0.008 | 0.033 | Median | 4841.8387 | 1001 | 528 |
| Lymphocytes % | Composition | lymphocyte /total cells | % | 0.81 | 0.68–0.96 | 0.016 | 0.057 | Median | 10.1541 | 1001 | 528 |
| Immune infiltration score | Established | immune cells /total cells | fraction | 0.81 | 0.68–0.96 | 0.017 | 0.058 | Median | 0.2823 | 1001 | 528 |
| Carcinoma delineation | Established | largest_perimeter_carcinoma /absolute_area_carcinoma | mm$^{-1}$ | 0.82 | 0.69–0.97 | 0.021 | 0.065 | Median | 0.0125 | 1001 | 528 |
| Intratumoral TIL % | Established | cell_percentage_lymphocyte_carcinoma | % | 0.83 | 0.70–0.99 | 0.039 | 0.095 | Median | 1.672 | 1001 | 528 |
| Plasma cell density | Composition | dens(plasma) | cells/mm$^2$ | 0.83 | 0.70–0.98 | 0.031 | 0.085 | Median | 109.7638 | 1001 | 528 |
| Carcinoma–lymphocyte adjacency | Novel spatial | fraction of neighbour-graph (delaunay) edges linking carcinoma to lymphocytes | fraction | 0.83 | 0.70–0.98 | 0.031 | 0.085 | Median | 0.0197 | 1001 | 528 |
| Lympho/macrophage ratio | Established | dens(lymphocyte)/dens(macrophage) | ratio | 0.85 | 0.72–1.01 | 0.062 | 0.137 | Median | 1.2653 | 1001 | 528 |
| Macrophage density | Established | dens(macrophage) | cells/mm$^2$ | 0.86 | 0.72–1.02 | 0.079 | 0.166 | Median | 381.45 | 1001 | 528 |
| Granulocyte density | Established | dens(granulocyte) | cells/mm$^2$ | 0.87 | 0.73–1.03 | 0.116 | 0.232 | Median | 121.645 | 1001 | 528 |
| Lymphocytes/mm$^2$ (whole tumor) | Established | dens(lymphocyte) | cells/mm$^2$ | 0.88 | 0.74–1.05 | 0.152 | 0.281 | Median | 400.95 | 1001 | 528 |
| TIL/tumor ratio | Established | dens(lymphocyte)/dens(carcinoma) | ratio | 0.88 | 0.74–1.05 | 0.153 | 0.281 | Median | 0.3357 | 1001 | 528 |
| CAF/tumor ratio | Established | dens(fibroblast in carcinoma)/dens(carcinoma) | ratio | 0.89 | 0.75–1.05 | 0.175 | 0.296 | Median | 0.0724 | 1001 | 528 |
| Tumor area % | Established | relative_area_carcinoma | % | 0.89 | 0.75–1.06 | 0.190 | 0.310 | Median | 37.9113 | 1001 | 528 |
| Granulocytes % | Composition | granulocyte /total cells | % | 0.90 | 0.76–1.07 | 0.225 | 0.353 | Median | 2.9308 | 1001 | 528 |
| Total immune density | Established | sum immune-cell densities | cells/mm$^2$ | 0.91 | 0.77–1.08 | 0.275 | 0.404 | Median | 1191.8332 | 1001 | 528 |
| Epithelial % | Composition | relative_area_epithelial_tissue (non-tumour epithelium /tissue area) | % | 0.91 | 0.76–1.08 | 0.262 | 0.398 | Gaussian mixture | 0.0134 | 1001 | 528 |

*Continued on next page*

*Supplementary Table 2 (continued)*

| feature | category | formula | unit | HR | CI_95 | p_value | q_value_BH | split_rule | split_cutoff | n | events |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Immune exclusion index | Established | dens(lymph in stroma)/dens(lymph in carcinoma) | ratio | 0.92 | 0.77–1.09 | 0.320 | 0.440 | Median | 12.4736 | 1001 | 528 |
| Tumor–stroma ratio | Established | carcinoma/(carcinoma+stroma) area | fraction | 0.92 | 0.78–1.09 | 0.349 | 0.463 | Median | 0.4818 | 1001 | 528 |
| Fibroblast density | Established | dens(fibroblast) | cells/mm$^2$ | 0.92 | 0.78–1.09 | 0.358 | 0.463 | Median | 770.3 | 1001 | 528 |
| Macrophage/tumor | Established | dens(macrophage)/dens(carcinoma) | ratio | 0.93 | 0.78–1.10 | 0.385 | 0.483 | Median | 0.2987 | 1001 | 528 |
| Macrophages % | Composition | macrophage /total cells | % | 0.93 | 0.79–1.11 | 0.423 | 0.517 | Median | 8.851 | 1001 | 528 |
| Fibroblasts % | Composition | fibroblast /total cells | % | 0.94 | 0.79–1.11 | 0.452 | 0.537 | Median | 18.2611 | 1001 | 528 |
| Normal tissue % | Established | (epithelial+other+vessel)/tissue area | % | 0.94 | 0.79–1.12 | 0.491 | 0.568 | Gaussian mixture | 4.2488 | 1001 | 528 |
| Granulocyte/tumor | Established | dens(granulocyte)/dens(carcinoma) | ratio | 0.95 | 0.80–1.12 | 0.528 | 0.596 | Median | 0.1149 | 1001 | 528 |
| Stroma area % | Established | relative_area_stroma | % | 0.95 | 0.80–1.13 | 0.581 | 0.639 | Median | 39.2449 | 1001 | 528 |
| Vascularization index | Established | dens(endothelial) | cells/mm$^2$ | 0.96 | 0.81–1.14 | 0.652 | 0.700 | Median | 68.8975 | 1001 | 528 |
| Vessel % | Composition | relative_area_vessel (vessel /tissue area) | % | 0.98 | 0.82–1.16 | 0.804 | 0.823 | Median | 1.245 | 1001 | 528 |
| Endothelial % | Composition | endothelial /total cells | % | 1.01 | 0.85–1.20 | 0.879 | 0.879 | Median | 1.6395 | 1001 | 528 |
| Blood % | Composition | relative_area_blood (blood /tissue area) | % | 1.04 | 0.87–1.23 | 0.696 | 0.729 | Gaussian mixture | 0.6527 | 1001 | 528 |
| Tumor cell density | Established | dens(carcinoma) | cells/mm$^2$ | 1.09 | 0.92–1.30 | 0.319 | 0.440 | Median | 1537.925 | 1001 | 528 |
| Interface immunity (20µm) | Established | interface lymphocyte density /bulk lymphocyte density | ratio | 1.13 | 0.95–1.34 | 0.174 | 0.296 | Median | 1.0763 | 1001 | 528 |
| Carcinoma % | Composition | carcinoma /total cells | % | 1.19 | 1.00–1.41 | 0.049 | 0.113 | Median | 37.698 | 1001 | 528 |
| Higher carcinoma cell–lymphocyte distance | Novel spatial | median distance (µm) from each carcinoma cell within the carcinoma compartment to its nearest lymphocyte (larger = immune exclusion) | µm | 1.20 | 1.02–1.43 | 0.033 | 0.085 | Median | 48.0712 | 1001 | 528 |
| Necrosis % | Composition | relative_area_necrosis (necrosis / segmented tissue area) | % | 1.27 | 1.07–1.52 | 0.006 | 0.029 | Gaussian mixture | 2.5339 | 1001 | 528 |
| Higher lymphocyte dispersion | Novel spatial | median nearest-neighbour distance (µm) between lymphocytes (larger = scattered, not clustered) | µm | 1.33 | 1.12–1.58 | 0.001 | 0.009 | Median | 13.1388 | 1001 | 528 |
| Higher endothelial cell–lymphocyte distance | Novel spatial | median distance (µm) from each endothelial cell to its nearest lymphocyte (larger = immune excluded from vasculature) | µm | 1.33 | 1.12–1.58 | 0.001 | 0.009 | Median | 23.9685 | 1001 | 528 |
| Tissue NLR | Established | dens(granulocyte)/dens(lymphocyte) | ratio | 1.37 | 1.16–1.63 | <0.001 | 0.009 | Median | 0.3378 | 1001 | 528 |
| UICC stage (per step) | Clinical anchor (UICC) | — | — | 1.42 | 1.31–1.54 | <0.001 | — | — (per stage step) | — | 1001 | 528 |

# B Report Evaluation

## Grading Guide

The reports are generated automatically from model outputs. Your review is the check that ensures every statement in a report is accurate, properly grounded in the underlying data, and clinically complete.

For grading you first need to break the report into individual statements that you then verify (**Claim-based Evaluation**). For each statement you check whether it is faithful to the model output it refers to, and whether any clinical inference drawn from it is correct.

You then step back and assess the report as a whole (**Report-level Evaluation**). You check whether it is internally consistent, and whether all clinically relevant findings are surfaced where a reader would expect them.

Whenever an error occurs, you will also rate its potential clinical harm, so that errors can be prioritized by how much they could affect patient care.

### How to review a report

1. Divide the report into discrete, atomic statements.
2. Grade each statement using Table 1: Claim-based Evaluation.
3. Grade the clinical harm (i.e. extent and likelihood) of the errors that occurred using Table 2: Clinical Harm Evaluation
4. Grade the report as a whole using Table 3: Report-level Evaluation.

### Scoring Category Definitions: Groundedness

- **Grounded** = The text is fully faithful to the module output, correctly preserving both direction and magnitude.
- **Partially** = The statement follows the right direction but is over- or understated relative to the data (e.g., output is 35%, but the report describes it as "extensive").
- **Ungrounded** = There is no underlying module output to support the claim, or the wording explicitly contradicts the data (e.g., output is 8%, but the report describes it as "moderate–high").

**Table 1: Claim-based Evaluation**

| Criteria | Description | Categories | Examples |
|---|---|---|---|
| Groundedness in Model Outputs | Applies to text that directly refers to module outputs in the report. Are these outputs correctly referenced and is the text faithful to the module | Grounded/Partially/ Ungrounded | • High Cellularity of 99.2%<br>• 3.2% Fibroblasts - low fibroblast content |

| Criteria | Description | Categories | Examples |
|---|---|---|---|
| | output? Does the wording preserve the output direction and rough magnitude? | | |
| Interpretation Correctness | Applies to claims that are the result of non-trivial logical reasoning from the module outputs. These claims often require clinical background knowledge and reasoning. Is the qualitative label, threshold bin, or clinical inference that the model suggests correct? | Correct/Incorrect/NA | • Trop2 H-Score threshold exceeds 200 and suggests potential eligibility for Trop-2-directed ADC therapy |
| Citation Correctness | Does the citation support the statement it refers to? | No / Partially / Yes | • Claim references a PubMed ID, but the cited paper does not substantiate the claim. |

**Table 2: Clinical Harm Evaluation**

| Criteria | Description | Categories | Examples |
|---|---|---|---|
| Extent | Anchored to clinical action. What is the extent of possible harm? | None / Mild or moderate / Severe or death | • Cellularity 99.2% vs 97.8% (None)<br>• Fibroblast % overstated prompting unnecessary confirmatory stain (Mild)<br>• Row-level misinterpretations that are not reflected in the overall subtyping and first-line |

| Criteria | Description | Categories | Examples |
|---|---|---|---|
| | | | therapy recommendation.( Moderate)<br>• Trop2 H-score wrong call leading to wrong first line therapy line (Severe or death) |
| Likelihood | Anchored to the workflow. What is the likelihood of possible harm? | Low / Medium / High | • Error is obvious or self-contradicted (Low)<br>• Plausible but routine review catches it (Medium)<br>• Plausible, decision-driving, no reliable catch (High) |

**Table 3: Report-level Consistency Checks**

| Criteria | Description | Categories | Examples |
|---|---|---|---|
| Omission | A clinically relevant finding that is obvious from the reported numbers is not highlighted in any of the summary sections from the report. | Present/Absent | • Tumor diagnosis<br>• Treatment Recommendation |
| Internal Contradiction | The overall assessment of the report contradicts the information given in the individual sections | Present/Absent | |

# C Report Prompt

```
You are a clinical pathology expert specializing in lung cancer diagnostics.
All cases are primary lung carcinomas.

Diagnostic Marker Section Content:
Interpret the provided markers and suggest a possible diagnosis based on the typical marker profiles of lung cancer subtypes. Do not provide any therapeutic interpretation or molecular testing recommendations in this section. Focus solely on subtype classification based on the marker data. For diagnostic markers, report each marker simply as positive or negative. Do NOT use the terms "TPS" or "H-score" and do not name the scoring method anywhere in this section's row interpretations or summary.

Diagnostic Marker Section context:
All marker values above 10% are positive. ADC: CK7 positive + p40 negative. When TTF1 negative then report TTF1 negative ADC, otherwise report TTF1 positive ADC. LCNEC: CK7 negative + synaptophysin/chromogranin A positive.
SqCC: CK5/6 positive + p40 positive + TTF1 negative

Prognostic Marker Section Content:
Interpret the provided prognostics markers in the context of therapeutic implications for lung cancer treatment. Do not make a subtype classification at this stage. Keep your assessment general about therapeutic implication based on the marker values and known thresholds.

Prognostic Marker Section context:
PD-L1 = [50-100%] -> recommend immune as first line therapy. (Monotherapy)
PD-L1 >0 until 50% -> Immune + chemotherapy. Suggest a specific immune checkpoint inhibitor if possible (e.g. pembrolizumab, atezolizumab). cMET: H-score: membranous H-score > 200 associated with improved response to cMET-directed ADCs (clinical trial context) Trop-2: membranous H-score >= 200 may be associated with improved response to TROP-2-directed ADCs (clinical trial context)

Molecular Panel Section Content:
For each row: one short sentence stating whether the alteration is actionable and any key threshold.

Molecular Panel Section context:
The molecular panel is the nNGM sequencing panel. Report only genes with detected mutations. Actionable alterations in lung cancer: KRAS G12C (sotorasib/adagrasib), EGFR exon 19 del / L858R (osimertinib),
BRAF V600E (dabrafenib+trametinib), ALK/ROS1/RET fusions (targeted inhibitors), MET exon 14 skip (capmatinib/tepotinib),
ERBB2 (trastuzumab deruxtecan), NTRK fusions (larotrectinib/entrectinib).
TP53 and KEAP1 are frequently mutated but not directly actionable; note them briefly. For the allele frequency: >20% is clonal (likely driver), <10% may be subclonal or artefact.

Section summary: 1-2 sentences on the most actionable finding(s) and molecular testing implications.

Guidelines:
- Use American English spelling throughout (tumor not tumour, vascularization, favor, characterize).
- For each row: write ONE short sentence (roughly 15 words or fewer) stating the key clinical fact.
  Do NOT use dashes (no em dash, no en dash, and no spaced hyphen used as a connector) and do NOT use
  semicolons. Commas are fine. Keep marker names intact (PD-L1, TTF-1, CK5/6, H-score).
  Reference thresholds where critical (PD-L1 >= 50%, MET H-score >= 200, TROP-2 eligibility).
  Example: "Expression is below the threshold for first-line pembrolizumab."
- For the section summary: 1-2 sentences maximum. Synthesize only the most actionable findings.
- Be factual. Do not fabricate data.
- For each row assign a traffic-light color (CSS hex string, or null) reflecting the PROGNOSTIC /
  PREDICTIVE value of the finding for the patient. The rule depends on the type of metric:

  Therapeutic / predictive markers (PD-L1, MET, TROP-2), GREEN or AMBER only, NEVER red:
    "#22a06b"  green: above threshold, a therapeutic option is indicated
                      (PD-L1 >= 50%, MET H-score >= 200, TROP-2 H-score >= 200). Opportunity wins:
                      color green even if high expression carries a worse baseline prognosis.
    "#e6a817"  amber: below threshold, no therapeutic option indicated (this is neutral, not red).

  Tissue Composition / TME Analysis / Immune Infiltration rows, do NOT choose a color:
    Their color is assigned deterministically in code from a 1001-case reference cohort and will
    OVERRIDE whatever you return, so return null for these rows. Each such row is annotated with its
    cohort position in brackets, e.g. "[cohort P72, high tertile, favorable end]" or
    "[cohort P50, middle tertile, descriptive]".
    Write in the register of a pathology report: ONE clinical sentence, roughly 20 to 30 words,
    covering: (1) briefly what the measure captures, (2) what it generally indicates prognostically in
    NSCLC, use the annotation's stated general direction ("higher values generally favorable" or
    "higher values generally unfavorable"), and state this for EVERY directional row INCLUDING ones
    where this case is intermediate, and (3) a qualitative read of THIS case's value (for example high,
    low, intermediate, sparse, abundant). Where it fits naturally and briefly, you MAY add a short
    clause on WHY the direction holds (a one-clause mechanism), but do not force it.
    Example: "Intratumoral TILs reflect cytotoxic lymphocytes engaging the tumor, and higher levels
```

```
    generally favor outcome because they signal active antitumor immunity, abundant in this case."
    Do NOT state the cohort tertile or percentile position in words (avoid "upper tertile", "low
    tertile", "mid cohort", "top of the cohort", "Nth percentile"): the bar already shows where the
    value sits. Do not use the word "descriptive" in the text.
    For metrics whose annotation says 'no established prognostic direction', give only what the measure
    captures and a qualitative read of the value, with NO favorable or unfavorable claim. Omit any prognostic
    ↪  comment.
    Vary the wording across rows so they do not read as a filled-in template. In particular, do NOT
    reuse one fixed phrase for the prognostic consequence: rotate how you express it rather than ending
    most rows with "a generally favorable/unfavorable feature in NSCLC". Draw on varied constructions
    such as "tends to portend better outcome", "usually carries a poorer prognosis", "a favorable
    prognostic sign", "often linked to reduced survival", "generally adverse in lung cancer" (do not
    reuse a single one). Stay consistent with the annotation (a value flagged low must read as low, not
    high). Do not copy the bracket punctuation, and follow the no-dashes and no-semicolons rule above.
    Rows with no annotation carry no cohort reference, so state the finding plainly.

  Cellularity (sample-adequacy metric):
    green >= 20% (adequate for molecular testing), amber 10-19% (borderline), red < 10% (insufficient).

  null (grey), NOT APPLICABLE: the metric carries no prognostic/predictive meaning. Use null for
    diagnostic lineage markers (TTF-1, p40, CK7, CK5/6, Chromogranin A, Synaptophysin) and for purely
    descriptive counts. These render grey.

  Red must only appear for a genuinely poor prognostic outcome or insufficient cellularity, never
  merely because a targeted therapy is unavailable.
```