nature cancer

Technical Report

https://doi.org/10.1038/s43018-023-00694-w

The artificial intelligence-based model ANORAK improves histopathological grading of lung adenocarcinoma

Received: 6 December 2022

Accepted: 21 November 2023

Published online: 10 January 2024

Check for updates

Xiaoxi Pan[1,2,19], Khalid AbdulJabbar[1,2,108], Jose Coelho-Lima[3,4,108], Anca-Ioana Grapa[1,2], Hanyun Zhang[1,2], Alvin Ho Kwan Cheung[5], Juvenal Baena[6,20], Takahiro Karasaki[5,7], Claire Rachel Wilson[6,8], Marco Sereno[9], Selvaraju Veeriah[5,7], Sarah J. Aitken[3,4], Allan Hackshaw[10], Andrew G. Nicholson[11,12], Mariam Jamal-Hanjani[7,13,14], TRACERx Consortium*, Charles Swanton[5,7,14], Yinyin Yuan[1,2,19,109] ✉, John Le Quesne[15,16,17,109] ✉ & David A. Moore[5,7,18,109] ✉

The introduction of the International Association for the Study of Lung Cancer grading system has furthered interest in histopathological grading for risk stratification in lung adenocarcinoma. Complex morphology and high intratumoral heterogeneity present challenges to pathologists, prompting the development of artificial intelligence (AI) methods. Here we developed ANORAK (pyrAmid pooliNg crOss stReam Attention networK), encoding multiresolution inputs with an attention mechanism, to delineate growth patterns from hematoxylin and eosin-stained slides. In 1,372 lung adenocarcinomas across four independent cohorts, AI-based grading was prognostic of disease-free survival, and further assisted pathologists by consistently improving prognostication in stage I tumors. Tumors with discrepant patterns between AI and pathologists had notably higher intratumoral heterogeneity. Furthermore, ANORAK facilitates the morphological and spatial assessment of the acinar pattern, capturing acinus variations with pattern transition. Collectively, our AI method enabled the precision quantification and morphology investigation of growth patterns, reflecting intratumoral histological transitions in lung adenocarcinoma.

Lung adenocarcinoma (LUAD), the most common type of non-small cell lung cancer, is histologically characterized by distinct growth patterns: lepidic, papillary, acinar, cribriform, micropapillary and solid[1] (Extended Data Fig. 1a). The proposed International Association for the Study of Lung Cancer (IASLC) grading system, based on a combination of the predominant growth pattern and high-grade patterns (cribriform, micropapillary and solid) within individual tumors, is highly prognostic[2]. However, there is interobserver variability among pathologists due to the challenges of consistently defining, recognizing and quantifying the wide spectrum of growth patterns[3]. This variability particularly affects differentiating lepidic, papillary and acinar patterns[2,4], as well as the estimated proportion of high-grade patterns in non-high-grade pattern-predominant tumors[2,5]. Accurate quantification is challenging when there are multiple admixed growth patterns across several histological sections, as is the case in most LUADs. This challenge is compounded by the difficulty of defining the

A full list of affiliations appears at the end of the paper. ✉e-mail: yyuan6@mdanderson.org; john.lequesne@glasgow.ac.uk; d.moore@ucl.ac.uk

cutoff between different patterns where they represent a spectrum of histological appearances[6]. This poses challenges for accurate prognostic inference and reproducibility in clinical studies.

Computer-assisted approaches powered by artificial intelligence (AI) have been widely applied to histological image analysis[7–11]. While some studies have applied deep learning models to LUAD growth pattern classification[12,13], automated IASLC grading by AI methods is yet to be explored. Moreover, previous deep learning methods were mainly based on patch-wise classification that predicts a histological subtype for each patch, overlooking the detailed morphological structure of patterns. To capture the distinct pattern morphology, we developed an AI method based on pixel-wise classification to segment growth pattern islands and automate the IASLC grading for risk stratification and outcome prediction.

In this study, we developed an AI method to segment LUAD growth patterns at the pixel level using hematoxylin and eosin (H&E) whole-slide images (WSIs) (Fig. 1a and Extended Data Fig. 1b,c) and applied it to 5,540 diagnostic slides from 1,372 cases, spanning four cohorts: TRAcking non-small cell lung Cancer Evolution through therapy (Rx) (TRACERx); Leicester Archival Thoracic Tumor Investigatory Cohort-Adenocarcinoma (LATTICe-A); The Cancer Genome Atlas (TCGA) LUAD; and Dartmouth Lung Cancer Histology Dataset (DHMC) (Fig. 1b). The growth pattern proportions, predominant pattern and IASLC grading of a tumor can be derived automatically based on growth pattern mapping (Fig. 1c). This pixel-wise segmentation method also revealed the morphological properties of growth patterns and enabled analysis of the degree of spatial heterogeneity, highlighting its advantages over patch-wise classification algorithms.

Results

A hierarchical AI model for growth pattern quantification

To spatially map complex growth patterns in LUAD, we developed ANORAK (pyrAmid pooliNg crOss stReam Attention networK), which encodes cross-stream interactions using a multi-order attention mechanism within convolutional neural networks[14] (Fig. 1a and Extended Data Fig. 1b,c). Moreover, a pyramid pooling module (PPM)[15] distributed global contextual information of growth patterns to guide high-level feature learning. ANORAK was trained on data annotated from 49 WSIs in the TRACERx 100 cohort (Extended Data Fig. 1a) by three thoracic subspeciality pathologists (Extended Data Fig. 1b), and validated on a total of 5,540 WSIs from 1,372 LUAD tumors across four cohorts (Fig. 1b and Table 1). This model enabled precision mapping of diverse growth patterns at pixel-level resolution, thereby facilitating automated grading and analysis of morphological intratumoral heterogeneity (Fig. 1c).

ANORAK generated promising outputs for growth pattern segmentation (Fig. 2a and Extended Data Figs. 2a,b and 3a). To validate the effectiveness of the developed model, we conducted the ablation study at the patch level (Extended Data Fig. 3b). Overall, multi-stream variants were more promising than single-stream ones, gaining an advantage by gathering different types of features. Moreover, methods with attention modules (multi-FO, multi-SO, ANORAK) achieved better overall performance, implying that the attention techniques came into effect. Specifically, first-order attention (multi-FO) improved performance by around 3% compared to the adding fashion (multi-ADD), while second-order attention (multi-SO) showed an approximate 5% improvement when compared to multi-FO. This suggested that high-level feature interactions across streams could be more effective than merging at low-level feature learning, highlighting the importance of high-level features in semantic segmentation[15,16]. The proposed model adopted both first-order and second-order attention modules, enhancing the overall performance with notable improvements. To compare this with existing methods, ANORAK outperformed several widely used approaches in semantic segmentation, including attention U-Net[17], DeepLabv3+ (ref. 18), DANet[19] and MedT[20], for growth pattern subtypes (0.4430–0.7463; Extended Data Fig. 3c) except for solid pattern (0.7170), which was lower than DeepLabv3+ (0.7381). ANORAK also achieved overall promising performance at the patch-level and WSI-level evaluations (patch-Dice: ANORAK: 0.6034, other methods: 0.3770–0.5691; WSI agreement: ANORAK: 60.00–65.31%, other methods: 16–48.98%; Extended Data Fig. 3c). Furthermore, the parameters of ANORAK are 4.10 million, that is, more lightweight than other convolutional models (6.67–15.55 million; Extended Data Fig. 3c). Taken together, the proposed model may have advantages in performance and computing over other methods.

In all four cohorts, AI-predicted growth pattern proportions were highly correlated with the pathologists' estimates (Fig. 2b and Supplementary Table 1), notably for the solid pattern (TRACERx 421, Spearman's rho = 0.79; LATTICe-A correlations against each pathologist's scoring, rho1 = 0.80, rho2 = 0.77, rho3 = 0.78; TCGA, rho = 0.67). The lowest correlations were observed for the micropapillary pattern (rho = 0.35–0.44 across three cohorts), which was also the pattern with the lowest interobserver agreement (LATTICe-A, 14.5–66.7%, average 39.8%; Extended Data Fig. 4a). When tumors were grouped according to their predominant pattern, the overall agreement rates between AI-predicted and manual scoring ranged between 50.18% and 67.96% (Supplementary Table 2 and Fig. 2c) across four cohorts. This is lower than the interobserver rates in LATTICe-A (53.49–74.08%; Extended Data Fig. 4b) but consistent with the known level of agreement between pathologists in previous studies (≥51.7%)[3,13]. The kappa statistics suggested a moderate agreement between AI and pathologists as well as inter-pathologists for predominant pattern assessment (averaged kappa index of AI-pathologist in four cohorts = 0.46; inter-AI-pathologists in LATTICe-A = 0.46; inter-pathologists in LATTICe-A = 0.49; Supplementary Table 3 and Extended Data Fig. 4b). Likewise, the overall agreement rates of AI-based grading according to the IASLC guidelines (AI grading hereafter) (65.73–76.80%; Supplementary Table 2 and Fig. 2e) were lower than the rates between pathologists in LATTICe-A (71.95–82.01%; Extended Data Fig. 4c,e), but the kappa statistics indicated a moderate agreement with manual grading, comparable with interobserver agreement (averaged kappa index of AI-pathologist in four cohorts = 0.47; inter-AI-pathologists in LATTICe-A = 0.50; inter-pathologists in LATTICe-A = 0.50; Supplementary Table 3 and Extended Data Fig. 4c). Interestingly, tumors with discrepant classification between AI and manual scoring had a notably higher intratumoral heterogeneity in growth pattern composition, measured using the Shannon diversity index based on pathological scores, compared to tumors concordant between AI and manual scoring (TRACERx 421, $P = 8.5 \times 10^{-7}$; LATTICe-A, $P1 < 2.22 \times 10^{-16}$, $P2 = 2.8 \times 10^{-12}$, $P3 < 2.22 \times 10^{-16}$; TCGA, $P = 0.00076$; Fig. 2d). A consistent trend was observed between discrepant and agreement classifications assessed by pathologists in LATTICe-A ($P < 2.22 \times 10^{-16}$, 4.3×10^{-13}, 1.6×10^{-15}; Extended Data Fig. 4d).

AI grading consistently improves patient risk stratification

Patients with IASLC grade 1 and 2 tumors as identified by AI had notably favorable disease-free survival (DFS) compared to patients with IASLC grade 3 tumors in TRACERx 421 ($n = 206$, $P = 0.003$, hazard ratio (HR) = 0.48, 95% confidence interval (CI) = 0.30–0.78) and LATTICe-A ($n = 729$, $P = 1.73 \times 10^{-7}$, HR = 0.53, 95% CI = 0.42–0.68; Fig. 3a). This prognostic effect remained notable when AI grading was incorporated in a multivariable model (TRACERx 421, $n = 206$, $P = 0.009$, HR = 0.51, 95% CI = 0.31–0.85; LATTICe-A, $n = 729$, $P = 0.001$, HR = 0.64, 95% CI = 0.49–0.84; Fig. 3b). The prognostic effect was slightly changed when tumor stage was replaced by tumor size (TRACERx 421, $P = 0.004$, HR = 0.48, 95% CI = 0.29–0.79; LATTICe-A, $P = 0.001$, HR = 0.64, 95% CI = 0.49–0.84; Extended Data Fig. 5a). The overall prognostic effect of the pair-wise comparison was consistently retained in the univariable (TRACERx 421, $P = 0.011$; LATTICe-A, $P = 7.81 \times 10^{-7}$) and multivariable analyses (TRACERx 421, tumor stage: $P = 0.033$, tumor size: $P = 0.014$; LATTICe-A, tumor stage: $P = 0.004$, tumor size: $P = 0.003$; Extended Data Fig. 5a).

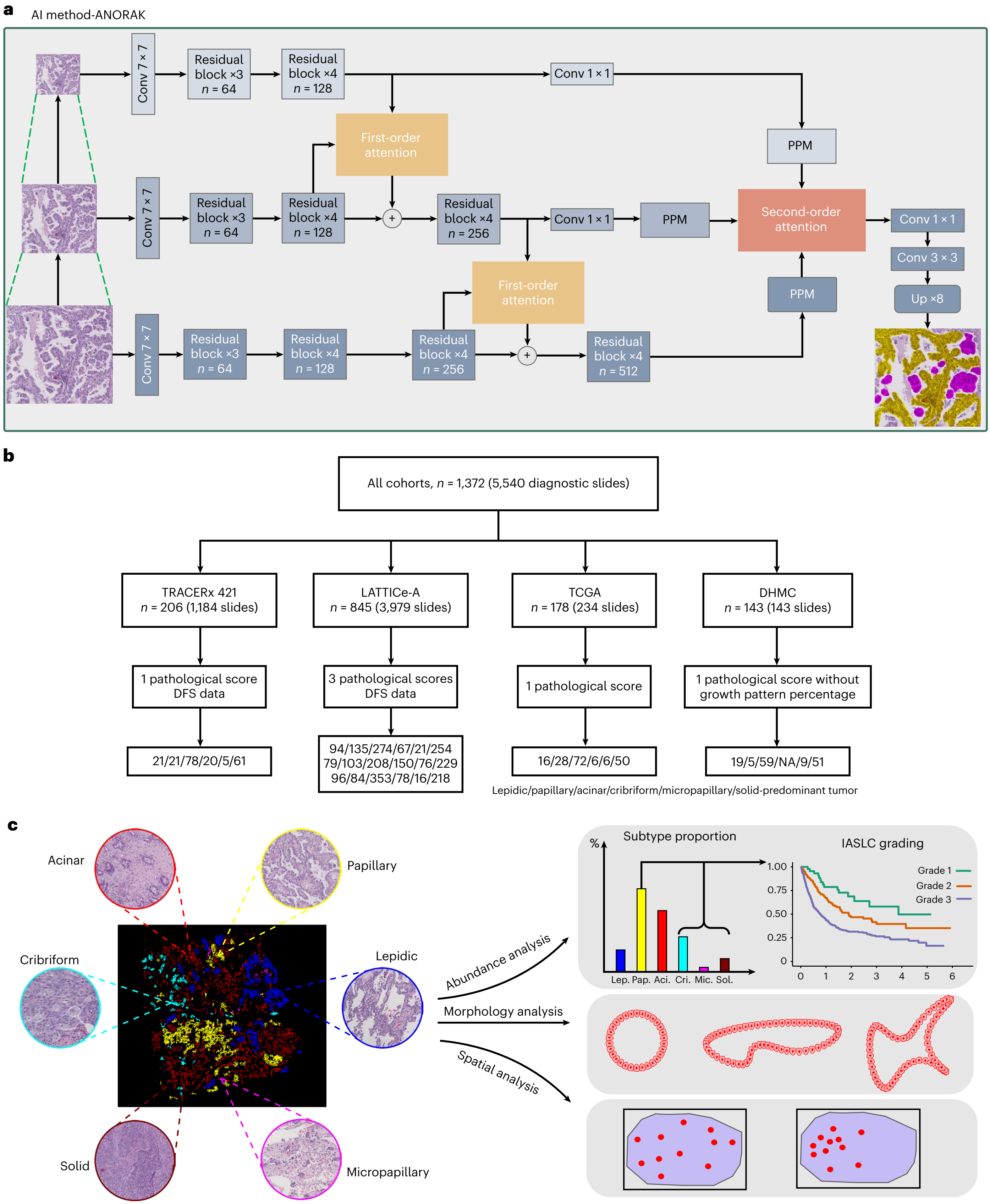


Fig. 1 | Proposed computational pipeline for precision mapping and spatial heterogeneity analyses. a, The deep learning network architecture for growth pattern segmentation integrating inputs over multiple spatial resolutions and delivering pixel-wise delineations. **b**, Overview of all the cohorts and available data. **c**, Downstream analyses enabled by the AI method, including abundance quantification, risk stratification and morphological and spatial heterogeneity analyses. PPM, Pyramid Pooling Module; Lep., lepidic; Pap., papillary; Aci., acinar; Cri., cribriform; Mic., micropapillary; Sol., solid; NA, not applicable.

Table 1 | Patient demographics (all cohorts)

Characteristic	TRACERx 421	LATTICe-A	TCGA LUAD	DHMC
Number of patients (diagnostic slides)	206 (1,184)	845 (3,979)	178 (234)	143 (143)
Pathological score available	1	3	1	1
Age, mean (minimum, maximum)	68.43 (37, 92)	67.66 (31, 86)	65.48 (42, 85)	–
Sex, *n* (%)				
Female	111 (53.88)	444 (52.54)	102 (57.30)	–
Male	95 (46.12)	401 (47.46)	76 (42.70)	–
Tumor stage, *n* (%)				
I	108 (52.43)	337 (39.88)	92 (51.69)	–
II	54 (26.21)	202 (23.91)	40 (22.47)	–
III	44 (21.36)	190 (22.49)	33 (18.54)	–
IV	0 (0)	0 (0)	12 (6.74)	–
Not applicable	0 (0)	116 (13.73)	1 (0.56)	–
Smoking status, *n* (%)				
Current smoker	88 (42.72)	259 (30.65)	–	–
Ex-smoker	101 (49.03)	419 (49.59)	–	–
Never smoker	17 (8.25)	64 (7.57)	–	–
Not applicable	–	103 (12.19)		
Adjuvant treatment, *n* (%)				
Yes	64 (31.07)	134 (15.86)	–	–
No	142 (68.93)	711 (84.14)	–	–
Type of surgery, *n* (%)				
Lobectomy or greater	180 (83.98)	640 (70.53)	–	–
Sublobar resection	26 (16.02)	89 (29.47)	–	–

To determine the prognostic information provided by AI compared to manual scoring and the clinical baseline characteristics, we focused on the large LATTICe-A cohort. While manual IASLC grading from all three pathologists was prognostic (Extended Data Fig. 5b–d), AI grading achieved a comparable performance with all three pathologists (Fig. 3b) in LATTICe-A. When Cox regression models were considered for predicting DFS (baseline; age, sex, tumor stage), AI grading (baseline + automated IASLC grading) and manual grading (baseline + a pathologist's manual IASLC grading), AI grading achieved a comparable performance with pathologists and clinical baseline for stage I–III tumors in LATTICe-A (*n* = 729, concordance index (C-index): AI = 0.682, 95% CI = 0.650–0.713; path 1 = 0.679, 95% CI = 0.645–0.713; path 2 = 0.680, 95% CI = 0.647–0.713; path 3 = 0.675, 95% CI = 0.644–0.707; baseline = 0.665, 95% CI = 0.633–0.697; Fig. 3c). Consistent performance was observed for stage I–III tumors in TRACERx 421 (*n* = 206, C-index: AI = 0.689, 95% CI = 0.625–0.752; path = 0.689, 95% CI = 0.625–0.752; baseline = 0.670, 95% CI = 0.608–0.733; Fig. 3c). In patients with early-stage tumors, the C-index of AI grading was comparable with pathologist grading but higher than baseline in TRACERx 421 (*n* = 108, C-index: AI = 0.700, 95% CI = 0.618–0.783; path = 0.695, 95% CI = 0.607–0.783; baseline = 0.665, 95% CI = 0.571–0.759; Fig. 3c). However, in LATTICe-A, the association between DFS and AI grading was consistently higher than the grading from pathologists (*n* = 337, C-index: AI = 0.643, 95% CI = 0.584–0.702; path 1 = 0.630, 95% CI = 0.570–0.690; path 2 = 0.615, 95% CI = 0.548–0.683; path 3 = 0.600, 95% CI = 0.526–0.673; baseline = 0.560, 95% CI = 0.495–0.625; Fig. 3c). Furthermore, once AI grading was added to manual grading (Supplementary Table 4), the prognostic value of the combined grading was consistently improved for stage I tumors (increment in C-index for path in TRACERx 421 = 0.013; path 1 = +0.023; path 2 = +0.028; path 3 = +0.043 in LATTICe-A; Fig. 3c), which was marginally higher than adding an additional manual grading in LATTICe-A (Extended Data Fig. 5e and Supplementary Table 5).

Taken together, these data suggest that AI grading adds independent prognostic value for patient stratification, particularly for stage I disease in which clinical decision-making regarding adjuvant therapy following surgery can be challenging in the absence of evidence for outcome benefit.

Assisting pathologists in challenging scenarios

To evaluate the utility of our AI method to assist pathologists with LUAD grading, we identified four specific scenarios and used the large LATTICe-A cohort with manual grading available from three pathologists. We focused on stage I LUAD tumors, a group of patients with an unmet need for predicting which patients are likely to relapse to guide early intervention, potentially with adjuvant therapy[21].

The first scenario consisted of cases with highly diversified growth patterns indicated by the Shannon diversity index (Fig. 4a), which was notably higher in cases with discrepant predominant patterns between AI and pathologists (Fig. 2d). When evaluated in cases with high growth pattern diversity based on the Shannon index derived from manual scoring, AI grading consistently obtained a higher C-index than pathological grading for DFS prediction (AI = 0.602, 95% CI = 0.485–0.720; path 1 = 0.590, 95% CI = 0.472–0.709, *n*1 = 169; AI = 0.602, 95% CI = 0.497–0.706; path 2 = 0.572, 95% CI = 0.453–0.692, *n*2 = 162; AI = 0.620, 95% CI = 0.537–0.704; path 3 = 0.578, 95% CI = 0.494–0.663, *n*3 = 167; stage I, Fig. 4a; stages I–III, Extended Data Fig. 6a; all models included baseline clinical parameters, same hereafter).

Second, we focused on tumors scored predominantly as lepidic or acinar by each pathologist, excluding any morphologically homogeneous tumor that received a score of 90% or more for either pattern[22]. There is an ongoing difficulty in the histopathological discrimination between in situ and invasive disease[4], and the distinction between invasive acinar and lepidic growth altered by interstitial fibrosis or iatrogenic compression with alveolar collapse can be particularly difficult. Differences in classification between pathologists can generate a shift between low and medium grade, which was observed among pathologists in the LATTICe-A cohort (Extended Data Fig. 4a). Therefore, these heterogeneously scored lepidic-predominant or acinar-predominant tumors present a challenging scenario to further test the added benefit of an AI grading system. AI grading consistently achieved a better performance in predicting DFS against pathological grading (AI = 0.658, 95% CI = 0.546–0.770; path 1 = 0.616, 95% CI = 0.513–0.718, *n*1 = 146; AI = 0.621, 95% CI = 0.530–0.711; path 2 = 0.587, 95% CI = 0.478–0.695, *n*2 = 136; AI = 0.703, 95% CI = 0.625–0.781; path 3 = 0.599, 95% CI = 0.512–0.687, *n*3 = 175; stage I, Fig. 4b; stages I–III, Extended Data Fig. 6b). There was a similar challenge in distinguishing between lepidic and papillary growth. When predominantly but heterogeneously presented (<90%) lepidic and papillary tumors were investigated in the context of comparing DFS prediction, AI grading consistently achieved a higher C-index (AI = 0.651, 95% CI = 0.420–0.882; path 1 = 0.619, 95% CI = 0.427–0.811, *n*1 = 92; AI = 0.658, 95% CI = 0.449–0.8670; path 2 = 0.614, 95% CI = 0.442–0.786, *n*2 = 77; AI = 0.602, 95% CI = 0.423–0.780; path 3 = 0.532, 95% CI = 0.373–0.692, *n*3 = 79; stage I, Fig. 4b; stages I–III, Extended Data Fig. 6b). The absence of statistical significance could be attributed to the relatively smaller number of patients and events in each group.

The third scenario was the detection of aggressive, high-grade patterns. Although there was a high concordance rate for cases composed predominantly of high-grade patterns (Extended Data Fig. 4e), the proposed IASLC grading system sets a 20% cutoff for high-grade patterns to qualify as grade 3, adding challenges to identify high-grade

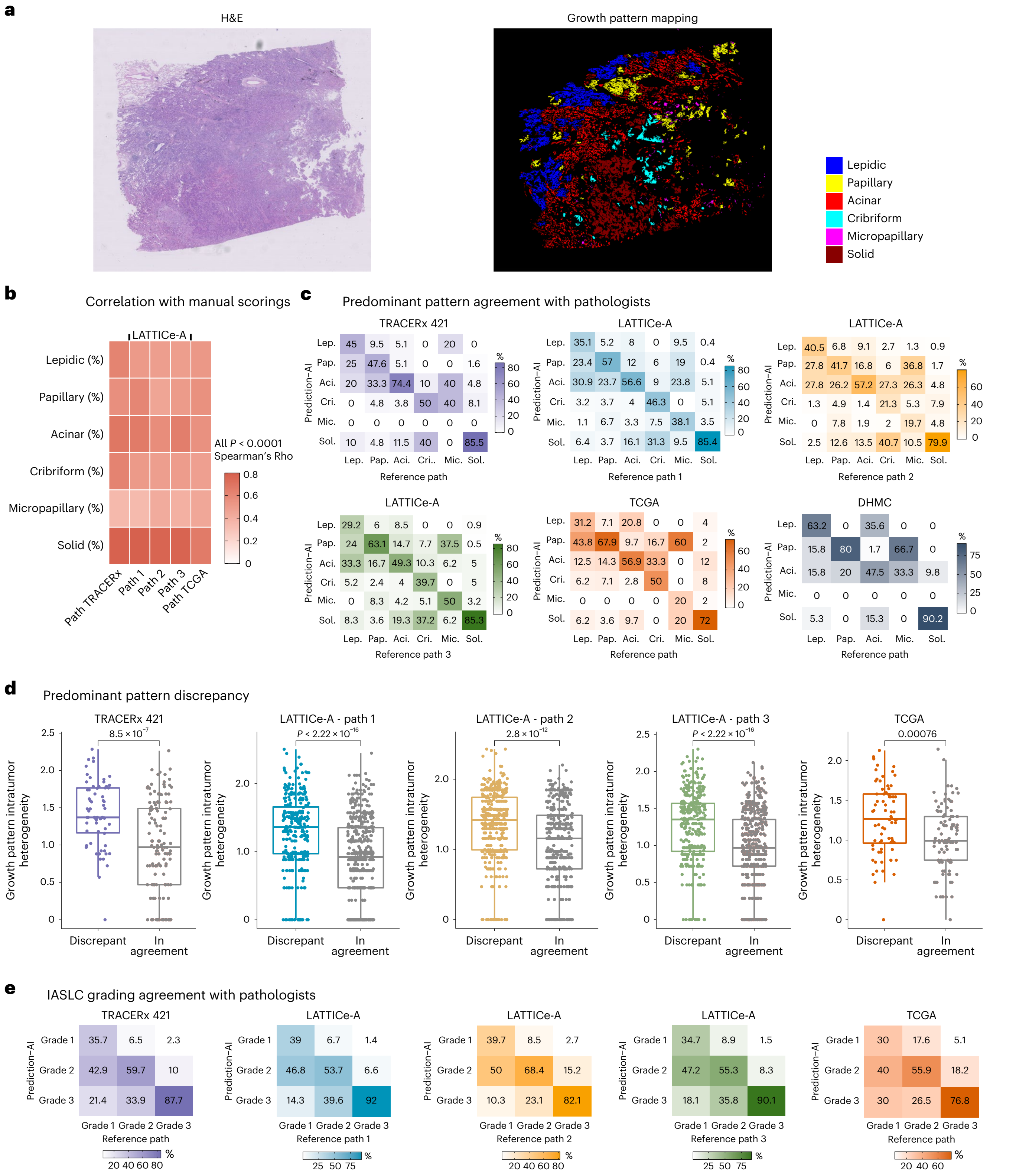


Fig. 2 | Performance of AI in the prediction and quantification of growth patterns. a, Segmentation example generated by ANORAK. **b**, Correlations of growth pattern proportions at the tumor level between AI and pathologists. Growth pattern proportions were not available in the DHMC cohort; thus, plots relevant to proportions were not illustrated for the DHMC (same in **d** and **e**). *P* values were corrected for multiple comparisons using the Benjamini–Hochberg method. **c**, Performance comparison with pathologists in predicting the predominant pattern per case (the cribriform predominant slide per tumor was not available in the DHMC cohort). **d**, Growth pattern intratumoral heterogeneity substantially contributed to the discrepancy between AI and pathologists (TRACERx 421, $P = 8.467 \times 10^{-7}$, $n = 206$; LATTICe-A, $P1 < 2.22 \times 10^{-16}$, $P2 = 2.816 \times 10^{-12}$, $P3 < 2.22 \times 10^{-16}$, $n = 845$; TCGA, $P = 0.0007632$, $n = 177$). Each *P* value was calculated using a two-sided Wilcoxon rank-sum test and not adjusted for multiple comparisons. The median value is indicated by a thick horizontal line; the first and third quartiles are represented by the box edges; the whiskers indicate 1.5× the interquartile range. **e**, Performance comparison with pathologists in the prediction of IASLC grading per case.

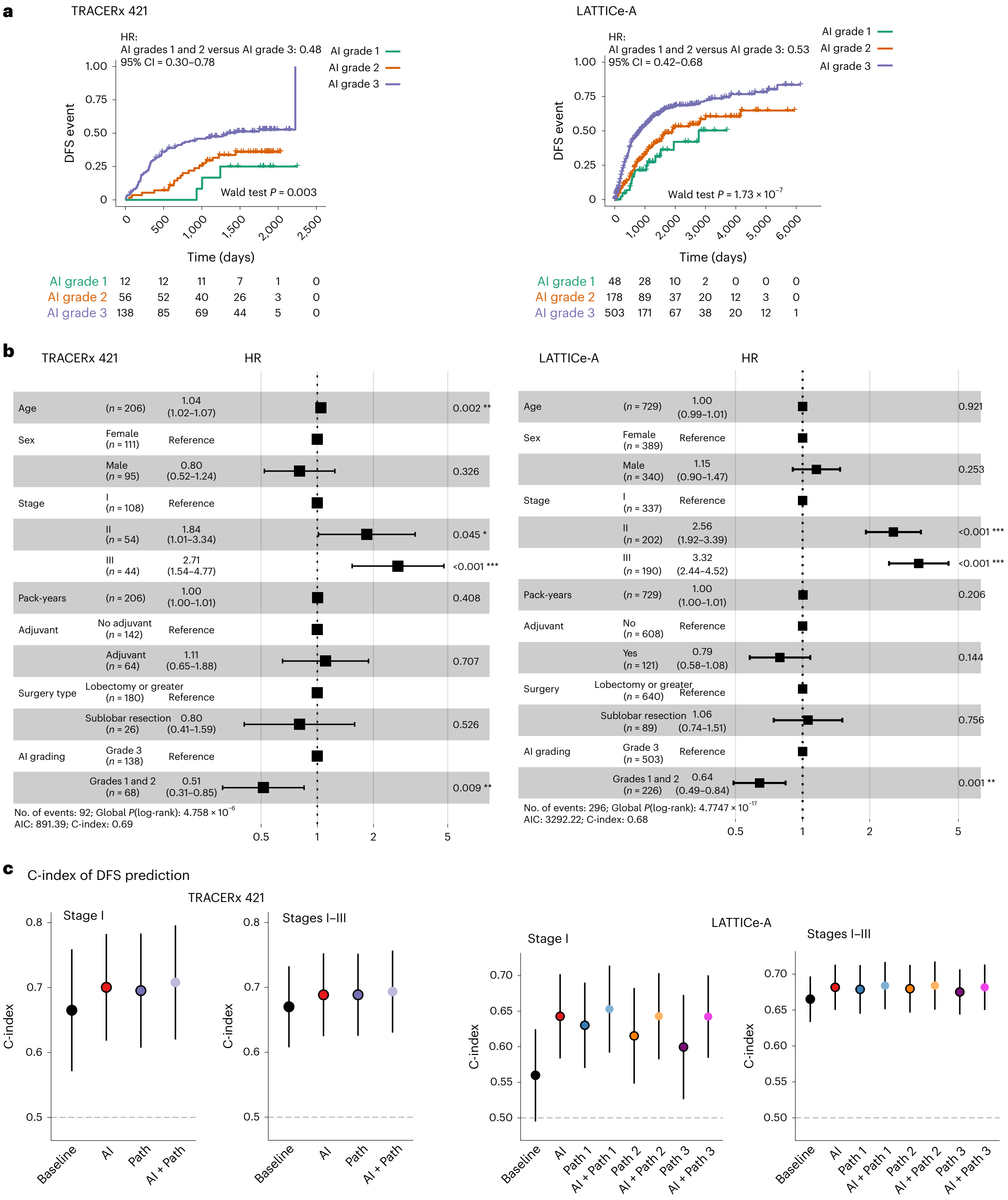


Fig. 3 | Survival analyses of AI and pathologist grading. a, Kaplan–Meier curves illustrating the difference in DFS according to AI grading. **b**, Multivariable Cox regression analyses showing that the prognostic effect of AI grading is independent of age, sex, tumor stage, smoking pack-years, adjuvant therapy and type of surgery (TRACERx 421: *P* = 0.009408, LATTICe-A: *P* = 0.00118). HRs of each variable with 95% CIs are shown on the horizontal axis; the *P* value was derived using a Wald test. **P* < 0.05, ***P* < 0.01, ****P* < 0.001. **c**, Comparison of DFS prediction measured according to C-index for stage I (TRACERx 421, *n* = 108; LATTICe-A, *n* = 337) and stage I–III (TRACERx 421, *n* = 206, LATTICe-A, *n* = 729) tumors, where the baseline characteristics included age, sex and tumor stage; AI included baseline parameters and AI grading; path included baseline parameters and pathologist grading; AI + path included baseline parameters, and AI and pathologist gradings. C-indexes with 95% CIs are shown on the vertical axis. AIC, Akaike information criterion.

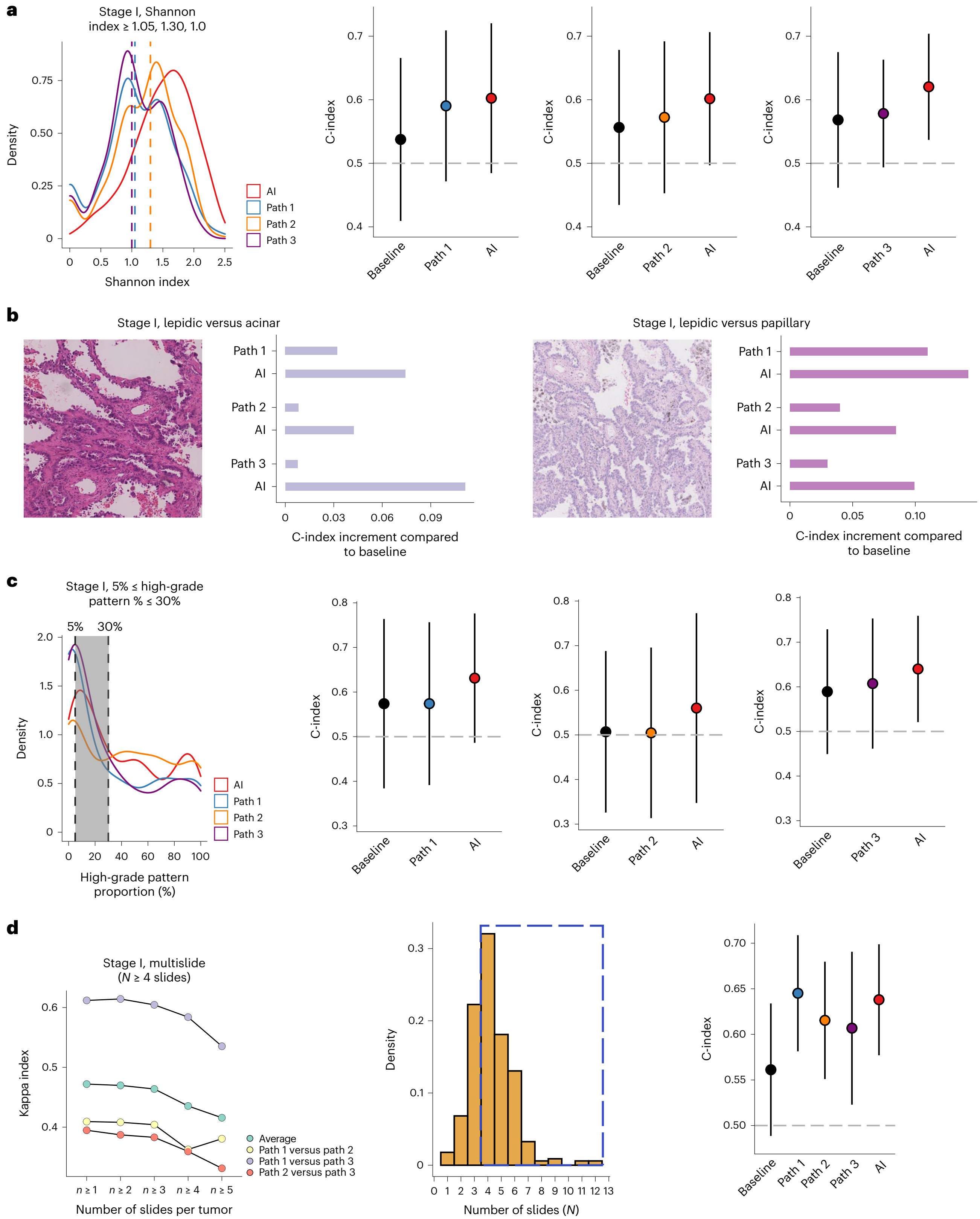


Fig. 4 | Assistance of AI in challenging scenarios for grading stage I tumors in LATTICe-A. a, Scenario 1: tumors with highly diversified growth patterns indicated by the Shannon diversity index (equal to or greater than the median). The vertical dashed lines indicate median values. Comparison of DFS prediction measured using the C-index (*n*1 = 169, *n*2 = 162, *n*3 = 167), where baseline included age and sex, AI included baseline parameters and AI grading, and path included baseline parameters and pathologist grading. C-indexes with 95% CIs are shown on the vertical axis (same with **c**,**d**). **b**, Scenario 2: differentiation between lepidic-predominant and acinar-predominant tumors (*n*1 = 146, *n*2 = 136, *n*3 = 175), and between lepidic-predominant and papillary-predominant tumors (*n*1 = 92, *n*2 = 77, *n*3 = 79). C-index improvement compared with baseline regarding DFS prediction. **c**, Scenario 3: tumors with high-grade patterns between 5% and 30% (*n*1 = 79, *n*2 = 63, *n*3 = 128, gray areas between two vertical dashed lines). **d**, Scenario 4: tumors with no fewer than four slides (*n* = 233, dashed box), for which the interobserver kappa index decreased.

patterns from non-high-grade pattern-predominant tumors. Therefore, we selected tumors with high-grade patterns (≥5%) at lower abundance (≤30%) as scored by each pathologist and compared their manual grading with AI grading. Such analyses allowed us to examine manually scored tumors, which may be 'close calls' among observers when determining the high-grade pattern cutoff. A higher C-index for AI grading was consistently observed compared with all pathologists' grading in predicting DFS (AI = 0.631, 95% CI = 0.486–0.776; path 1 = 0.574, 95% CI = 0.392–0.757, *n*1 = 79; AI = 0.560, 95% CI = 0.347–0.773; path 2 = 0.505, 95% CI = 0.313–0.696, *n*2 = 63; AI = 0.640, 95% CI = 0.521–0.759; path 3 = 0.607, 95% CI = 0.461–0.753, *n*3 = 128; stage I, Fig. 4c; stages I–III, Extended Data Fig. 6c).

Finally, we considered cases with high numbers of diagnostic slides per tumor (Fig. 4d), defined as four or more slides (*n* = 233, decreased kappa index in Fig. 4d). In these cases, AI grading achieved a C-index higher than average for the manual grading but lower than pathologist 1 in predicting DFS (AI = 0.638, 95% CI = 0.577–0.699; path 1 = 0.645, 95% CI = 0.581–0.709; path 2 = 0.615, 95% CI = 0.551–0.680; path 3 = 0.607, 95% CI = 0.523–0.691; stage I, Fig. 4d; stages I–III, Extended Data Fig. 6d).

These data indicated that our proposed AI method was not inferior to pathological grading and could assist pathologists to grade growth patterns in certain challenging scenarios.

Acinar morphology and spatial heterogeneity

Precise spatial delineations of growth patterns allowed us to study the spatial configuration of tumors as morphologically distinct pattern islands (Fig. 2a and Extended Data Figs. 2a,b and 3a). Acinar growth, often considered as an intermediate state during the transition of morphological patterns[6,23], was also the most prevalent pattern in stage I tumors in the LATTICe-A cohort (Fig. 5a). The area of individual acinar islands was similar to that of micropapillary islands, and smaller than those of other patterns (Fig. 5b). These data led us to investigate the importance of morphological features and spatial distribution of acinar islands that may be indicative of histology pattern transition.

We used area and shape measured using pixel number and solidity index (Extended Data Fig. 7a) to represent the morphological features of individual acinar islands. Acinar island area and shape were notably different in tumors (≥5% of acinar) with different predominant patterns (TRACERx 421 *n* = 173; LATTICe-A *n* = 654; Extended Data Fig. 7b). Smaller acinar islands were enriched in lepidic-predominant tumors compared to acinar-predominant and papillary-predominant tumors (TRACERx 421 *P* = 0.00052; LATTICe-A $P = 5.4 \times 10^{-12}$; Fig. 5c and Extended Data Fig. 7c). This may reflect the acinar structures in lepidic-predominant disease frequently representing airspaces with iatrogenic collapse[24]. The area of acinar islands in high-grade pattern-predominant (cribriform, micropapillary and solid) tumors were also smaller than those in acinar-predominant and papillary-predominant tumors (TRACERx 421 $P = 9.8 \times 10^{-11}$; LATTICe-A $P < 2.22 \times 10^{-16}$; Fig. 5c and Extended Data Fig. 7c). Notably, this area feature was a strong discriminator between acinar-predominant and cribriform-predominant tumors (TRACERx 421 *P* = 0.0007; LATTICe-A $P = 1.5 \times 10^{-7}$; Fig. 5d), indicating that acini may form differently in acinar-predominant tumors compared to others. The transition from an acinar to a cribriform pattern may frequently occur to large acinar islands through gland fusion (Extended Data Fig. 7e), while smaller acinar structures may remain. Alveolar architectures in airspace detected in acinar-predominant tumors might also be supporting large 'glands'. Acinar islands with regular shapes were enriched in high-grade-predominant tumors compared with lepidic subtypes (TRACERx 421 *P* = 0.0024; LATTICe-A $P = 4.1 \times 10^{-7}$; Fig. 5e and Extended Data Fig. 7d), which is again consistent with morphological variance due to the compressibility of lepidic growth. Taken together, the morphological features of acinar islands vary notably in tumors predominantly enriched with different patterns (Fig. 5f).

To investigate the spatial arrangement of acinar patterns, we developed an acinar scattering score that measured the degree of acinus dispersion. A low score indicated locally clustered acinar islands, while a high score implied a dispersion of acinar islands throughout the tissue (Extended Data Fig. 7f). Low acinar scattering was found more frequently in lepidic-predominant tumors compared to all others (TRACERx 421 *P* = 0.017; LATTICe-A *P* = 0.004; Fig. 5g), indicating that clustered acinar islands may reflect the compression induced by iatrogenic collapse and may also suggest that the transition from lepidic to acinar occurs in an organized manner[25]. We next explored acinar scattering in the context of outcome prediction. Tumors with highly scattered acini were associated with reduced DFS compared to lowly scattered tumors (TRACERx 421 *n* = 205, *P* = 0.003, HR = 1.89, 95% CI = 1.25–2.86; LATTICe-A *n* = 837, $P = 5.09 \times 10^{-7}$, HR = 1.63, 95% CI = 1.35–1.98; Fig. 5h) in univariate analysis. In a multivariable model incorporating acinar scattering and AI grading, acinar scattering was independent of AI grading (TRACERx 421 *P* = 0.004; LATTICe-A $P = 2.61 \times 10^{-5}$; Fig. 5i). These data suggest that acinar scattering may be a potential pattern reflecting histological transition events, and that high scattering may be a morphological phenotype indicating poor prognosis, which can be assessed from H&E images.

Discussion

We have developed an AI method ANORAK for the precise classification of growth patterns in LUAD. To the best of our knowledge, this is the first AI method to dissect LUAD growth patterns at the pixel level and be tested in over 1,000 cases, setting a benchmark in automated grading of LUAD. Our method can automatically estimate growth pattern proportions and predominant patterns within a tumor, providing an unbiased and automated pipeline for determining IASLC grading in LUAD. Moreover, the precise delineation of growth patterns can provide insights into the heterogeneous landscape of LUAD, which cannot be addressed by patch-wise classification methods.

The AI method was evaluated in four cohorts, comprising a total of 1,372 tumors. The overall agreement of predominant pattern at the tumor level between AI and pathologists across four cohorts was moderate, which is consistent with the inter-pathologist agreement in the LATTICe-A and DHMC cohorts[13]. Similar results were found in previous studies. Boland et al.[3] reported an agreement of 51.7% between two pathologists for a large cohort of individuals with LUAD (*n* = 534), while Thunnissen et al.[4] showed good agreement for typical cases and fair agreement for difficult cases by comparing scores from 26 pathologists. In addition, tumors with a discrepant predominant pattern classification between AI and manual scoring were more heterogeneous compared to tumors in agreement. Previous attempts were made to determine how clonal evolution is reflected in growth pattern heterogeneity through the identification of molecular alterations that accompany the transition between growth patterns[6]. This detailed analysis in a small number of tumors found that changes in expression, rather than mutations, accompanied the transition; as such, clear evidence of divergent tumor clones reflected in the growth pattern was not identified. On a larger scale, in the TRACERx study, although without specific focus on sampling to capture divergent growth patterns, there was a tendency for tumors to evolve from low-grade or mid-grade to higher grade growth patterns in individuals with LUAD where an ancestor–descendant relationship could be described based on clonal or subclonal loss of heterozygosity[22].

The proposed IASLC grading system was originally introduced to improve prognostication using tumor morphology[2]. In our study, AI grading improved the performance of predicting DFS compared to the baseline and pathological grading for stage I tumors, and be comparable for stage I–III tumors. Moreover, the prognostic value of AI grading was independent of clinical parameters in the TRACERx 421 and LATTICe-A cohorts. In typical clinical practice, the colineage of postsurgical recurrence is not definitively confirmed, although

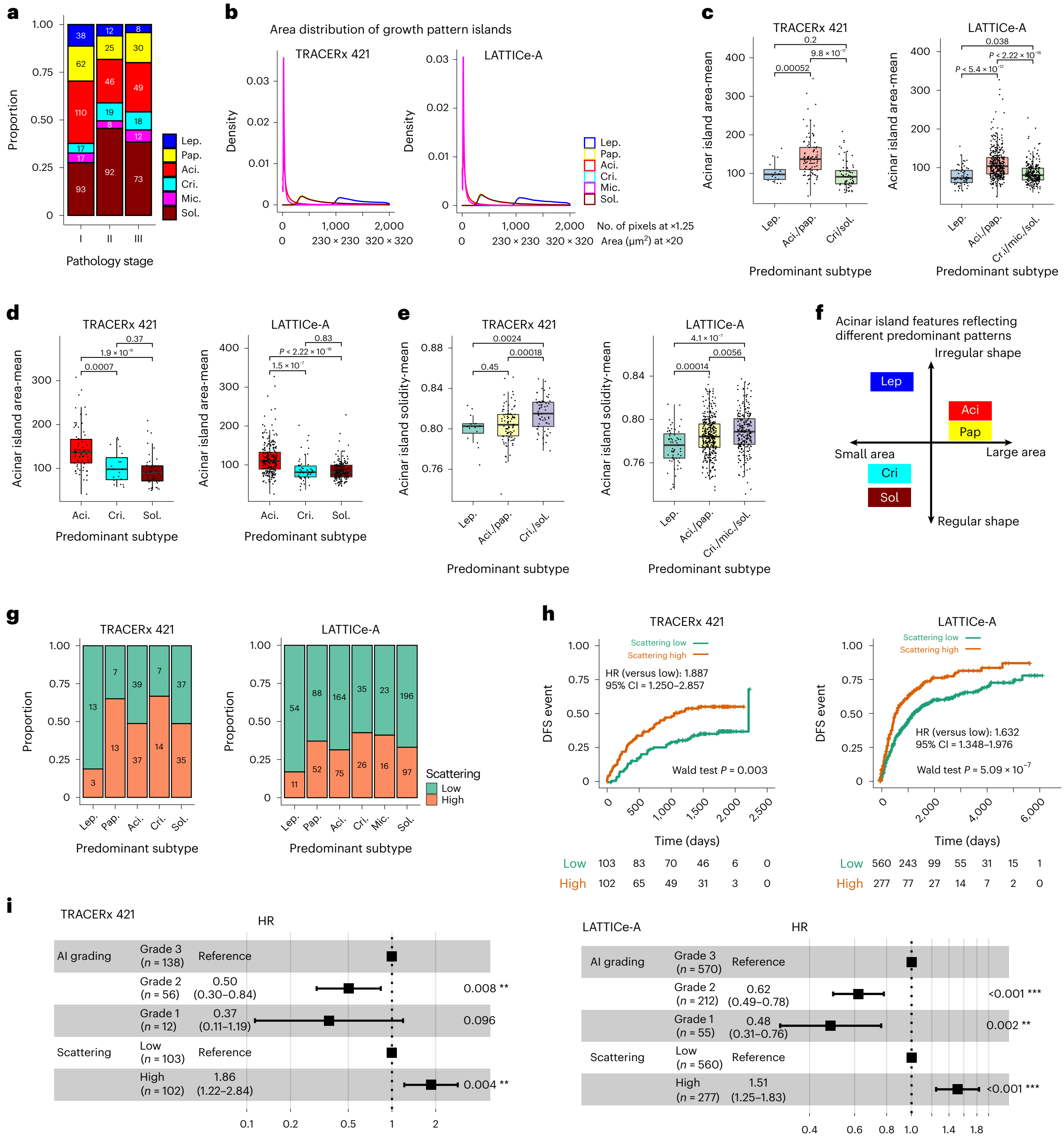


Fig. 5 | Characterization of tumors with acinar morphological features and spatial heterogeneity. a, LUAD subtype distribution across stages in LATTICe-A showing that acinar is the most prevalent pattern in stage I tumors. **b**, Area distribution of growth pattern islands delineated by AI in TRACERx 421 and LATTICe-A, indicating that the areas of acinar islands are similar to micropapillary islands, but smaller than lepidic, papillary, cribriform and solid islands. **c**, Smaller acinar island areas were enriched in lepidic-predominant (TRACERx 421, $P = 0.0005161$, $n = 108$; LATTICe-A, $P = 5.413 \times 10^{-12}$, $n = 420$) and high-grade-predominant tumors (TRACERx 421, $P = 9.797 \times 10^{-11}$, $n = 157$; LATTICe-A, $P < 2.2 \times 10^{-16}$, $n = 593$) compared to acinar-predominant and papillary-predominant tumors. **d**, Acinar island areas were notably smaller in cribriform-predominant tumors compared to acinar-predominant tumors (TRACERx 421, $P = 0.0006956$, $n = 95$; LATTICe-A, $P = 1.515 \times 10^{-7}$, $n = 290$). **e**, Acinar island shapes were notably regular in high-grade-predominant tumors compared to lepidic-predominant tumors (TRACERx 421, $P = 0.002439$, $n = 81$; LATTICe-A, $P = 4.118 \times 10^{-7}$, $n = 295$). **c**–**e**, Each point is a tumor; the *y* axis is the mean area (**c**,**d**) or solidity index (**e**) of all the individual acinar islands within a tumor. The *P* value was calculated using a two-sided Wilcoxon rank-sum test not adjusted for multiple comparisons. The median value is indicated by a thick horizontal line; the first and third quartiles are represented by the box edges; the whiskers indicate 1.5× the interquartile range. **f**, Acinar morphological features reflecting different growth patterns; small-area acinar islands with irregular shapes were more likely observed in lepidic-predominant tumors, whereas in cribriform-predominant and solid-predominant tumors, small-area acinar islands with a regular shape were enriched. **g**, Spatial arrangement of acinar islands across predominant subtypes. **h**, Kaplan–Meier curves comparing tumors with low and high levels of acinar scattering for TRACERx 421 and LATTICe-A. **i**, Multivariable Cox regression analyses showing that tumors exhibiting a high degree of acinar scattering were linked to decreased DFS compared to tumors with low acinar scattering, independent of AI grading in TRACERx 421 ($P = 0.004209$) and LATTICe-A ($P = 2.61 \times 10^{-5}$). HRs of each variable with 95% CIs are shown on the horizontal axis; the *P* value was derived using a Wald test. ${}^{**}P < 0.01$, ${}^{***}P < 0.001$.

data from the TRACERx 421 cohort showed that only two out of 49 cases of clinically classified postsurgical recurrence were of different lineage using whole-exome sequencing[26]. While we acknowledge that these uncommon events limit the ability to predict recurrence from resection specimens, this applies equally to both our method and established practices.

The LATTICe-A cohort, consisting of 845 tumors with scores from three pathologists, allowed a comprehensive investigation of the clinical impact of the AI method and showed its benefit as a morphological biomarker. This benefit was slightly higher than that brought by an additional manual grading for stage I tumors, and was comparable with additional manual grading for stage I–III tumors. Furthermore, analyses of manual scoring demonstrated that tumors with multiple slides and intratumoral morphological heterogeneity were particularly challenging cases. In these cases, AI grading achieved a stronger predictive ability compared to manual grading for stage I tumors. Because stage I patients frequently receive surgical resection without adjuvant therapy, the accurate prediction of recurrence, to better target individual patients for adjuvant therapies, is critical. These data illustrate the clinical utility of our AI method for stage I tumors, which could potentially be used as an alternative or independent variable to manual grading, or be applied specifically to challenging cases.

The AI method enables the spatial profiling of growth patterns at the pixel level, allowing morphological and spatial heterogeneity analyses at the growth pattern island level. This would be unattainable with alternative manual or patch-wise classification methods. We used the area and solidity index to measure acinar island morphology and found that small acinar islands were enriched in lepidic-predominant and high-grade-predominant tumors, while the shape of these small acini in lepidic-predominant tumors was more irregular than high-grade-predominant tumors. This may reflect tumor cell biological and microenvironmental differences regarding the formation of acinar structures within the context of different predominant architectures. Because acinar morphological features were obtained by averaging thousands of acinar islands within a tumor, noise due to island segmentation was mitigated (Supplementary Figs. 1–7). We also developed a metric for measuring the spatial distribution within the tissue space of acinar islands, termed acinar scattering. Low acinar scattering was notably associated with lepidic-predominant tumors compared to others, suggesting that acinar spatial distribution may reflect the transition of growth patterns toward more aggressive behavior. High acinar scattering was correlated to unfavorable outcomes, independent of AI grading.

This study has some limitations. The Dice coefficient of ANORAK is still limited, indicating that error modes exist. Intratumoral and tumor microenvironment heterogeneity may result in variations in growth pattern morphology, making segmentation more challenging, specifically among lepidic, papillary and acinar patterns. Meanwhile, the patching operation during the training and testing stages may limit the field of view, thus losing context information. Stain color shift may also have the potential for misclassification despite the color augmentations and normalizations applied to mitigate this impact. These factors may contribute to local error modes, which, when accumulated, may result in errors at the WSI level. In addition, because the model counted the number of pixels to determine the predominant pattern per tumor, and the area of micropapillary islands was smaller than the papillary structures[27], the discrepancy between AI and pathologists regarding papillary-predominant and micropapillary-predominant patterns may be considered another error mode. Furthermore, because we only collected histopathology annotations from invasive non-mucinous LUAD as training data, invasive mucinous and preinvasive tumors with distinct morphologies are therefore outside of the scope, which may generate inaccurate results or completely fail if applied to such samples. In addition, we selected a 'challenging case series' from the LATTICe-A cohort, because the other cohorts considered in this study had fewer cases satisfying the selection criteria. However, LATTICe-A is not a screening-based cohort. It is therefore crucial to validate the potential clinical benefits of AI grading in further cohorts that include screening-detected tumors. Because there are no other studies reporting the importance of acinar spatial arrangement, further validations and studies of the biological implications of acinar scattering are needed.

In summary, the AI method we developed can automate the predominant growth pattern and IASLC grading for LUAD tumors, achieving a moderate agreement with pathologists; this was validated in four cohorts consisting of 1,372 cases. In the TRACERx 421 and LATTICe-A cohort, AI grading was an independent prognostic indicator and had a stronger prognostic ability than pathological grading alone for stage I tumors in the LATTICe-A cohort. The prognostic performance of AI grading was further underlined in challenging scenarios consisting of cases with multiple slides and greater intratumoral heterogeneity. Furthermore, specific morphological features of tumor acini have the potential to infer different underlying tumor biology, with the spatial heterogeneity of acinar islands reflecting divergent tumor behavior and prognosis.

Methods

Study cohorts

TRACERx is a multi-center, prospective study, which began recruitment in April 2014 (https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent research ethics committee, ref. no. 13/LO/1546). Formalin-fixed paraffin-embedded and H&E-stained histopathology diagnostic slides were scanned using the NanoZoomer S210 digital slide scanner (catalog no. C13239-01) and NanoZoomer digital pathology system v.3.1.7 (Hamamatsu) at ×40 (0.228 µm per pixel resolution)[28,29]. LATTICe-A is a retrospective series of all consecutively resected primary LUAD tumors at a single UK surgical center between 1998 and 2014. The work was ethically approved by a UK National Health Service research ethics committee (ref. no. 14/EM/1159) and complies with Strengthening the Reporting of Observational Studies in Epidemiology guidelines. All archived slides containing tumor material were used to capture the full diversity of each lesion. Slides were dearchived and scanned using a Hamamatsu NanoZoomer XR at ×40 (0.226 µm per pixel resolution)[23,29]. Available diagnostic slides from the TCGA LUAD[30] were downloaded from https://portal.gdc.cancer.gov/ in 2021. The DHMC[13] was downloaded from https://bmirds.github.io/LungCancer/ in 2021. Further information on the research design is available in the Nature Research Reporting Summary linked to this article.

The training set of the AI method consisted of 49 WSIs from 49 patients in the TRACERx 100 cohort[28,29]. The WSIs were sparsely annotated by three independent thoracic subspeciality pathologists, yielding 3,662 patches (768 × 768 pixels at ×20, approximately 0.45 µm per pixel) of annotations for six typical growth patterns (Extended Data Fig. 1a) and non-tumor areas, for example, normal tissue and blank areas.

The AI method was then applied and evaluated on a total of 5,540 WSIs from four cohorts, which were collected, processed and scanned independently. This included patients with invasive non-mucinous LUAD as primary diagnosis (excluding adenocarcinoma in situ, minimally invasive adenocarcinomas and other variants) from the TRACERx 421 cohort ($n = 206$, 1,184 slides)[22,26], LATTICe-A cohort ($n = 845$, 3,979 slides)[23], TCGA LUAD cohort ($n = 178$, 234 slides)[30], DHMC cohort ($n = 143$, 143 slides)[13] (Table 1). TRACERx 100 is a subset of TRACERx 421. For the TRACERx 421 and LATTICe-A cohorts, slides were from all the diagnostic blocks containing tumor cells. For the DHMC cohort and most patients (91%) in the TCGA cohort, only one slide was available. Hence, we only considered these two cohorts for agreement performance comparison. No statistical method was used to predetermine sample size but our sample sizes are similar to those reported in previous publications[13,22,26,29,30] and subject to available diagnostic slides. Blinding and randomization were not relevant because this was an observational study. Patients were not allocated to any interventions and they were followed up and assessed as per routine practice. No results from this study were reported back to patients, so there is no likelihood of people changing their behaviors based on these findings. The deep learning model was trained without knowing the outcome of patients, which represents a form of blinding.

Manual pathological grading of growth patterns, as well as individual pattern proportion scoring, were available for the TRACERx 421, LATTICe-A and TCGA cohorts. The DHMC cohort only had predominant pattern data for each slide. In the LATTICe-A cohort, three independent consultant-level thoracic subspeciality pathologists provided growth pattern scoring for each tumor.

In the TRACERx 421 cohort, DFS was defined as the period from the date of registration to the time of radiological confirmation of the recurrence of the primary tumor registered for the TRACERx or the time of death by any cause. During the follow-up, three participants with LUAD (CRUK0512, CRUK0428 and CRUK0511) developed new primary cancer and subsequent recurrence from either the first primary lung cancer or the new primary cancer diagnosed during the follow-up. These cases were censored at the time of the diagnosis of the new primary cancer for DFS analysis because of the uncertainty of the origin of the third tumor[22].

In the LATTICE-A cohort, recurrence data were obtained from the examination of patient records, notably paper notes and radiological databases, to identify the date of radiologically or biopsy-confirmed recurrence. Cancer-specific death was determined by the presence of lung cancer in the cause of death in the death certificate. Overall survival refers to the date of death.

Deep learning model architecture

We developed a deep learning-based model[14] ANORAK which leveraged cross-stream interaction to recognize and segment six histological patterns (lepidic, acinar, papillary, micropapillary, cribriform and solid) on WSIs at the pixel level. The model applied ResNet50 (ref. 31) as the backbone with customized modifications to account for the limited training data. It encoded three streams (coarse, intermediate and fine) with different scales of information to gather abundant features at different resolutions (×10 at approximately 0.9 µm per pixel, ×5 and ×2.5). The first-order attention (Extended Data Fig. 1c) introduced global contextual information at an early stage to guide low-level feature learning and enable the first round of interactions between streams. Each output in the coarse and intermediate streams was then fed into a convolution layer to align the depth dimension with the fine stream output. A PPM[15] (Extended Data Fig. 1c) was used to integrate high-level features. Afterwards, such features were forwarded to a second-order attention module, learning the relationship of streams to extract more discriminative features, and driving high-level feature exchanging between streams (Extended Data Fig. 1c and Fig. 1a).

Implementation and evaluation

Before training, the annotated tiles were divided into nonoverlapping patches, except for patches at the bottom and right edges, with a size of 768 × 768 pixels at ×20. During training, four data augmentation strategies were used to mitigate overfitting: random rotation within 90 degrees; random width-shift and height-shift up to 20% of the input width and height; randomly zooming in or out in a range of (0.8, 1.2); and random adjustment of the saturation within (0.8, 2.0) and hue within (−0.1, 0.1). Color augmentation was not applied to the cross-validation stage because data were from the same cohort. The model was trained for 60 epochs with a batch size of eight. Cross-entropy loss was applied as the objective function, which was minimized by the Adam optimizer with a step-wise learning rate. The initialization rate was set to 10^{-3} for the first ten epochs; then, it was decreased by ten times for the next 40 epochs, which was then followed by another ten times of decreasing (10^{-5}) for the remaining ten epochs. The pipeline was implemented with Python v.3.8, tensorflow-gpu v.2.2, keras v.2.4.3, h5py v.2.10.0, numpy v.1.20.3, opencv-python v.4.5.3.56, pandas v.1.3.2, pillow v.8.3.1 and scipy v.1.7.1.

The ablation experiments at the patch level included comparisons with the baseline method (single-stream), multi-stream with the element-wise add combination (multi-ADD), multi-stream with first-order attention alone (multi-FO), multi-stream with second-order attention alone (multi-SO) and the proposed ANORAK model (multi-FO and multi-SO). The proposed model was compared against other widely used approaches in semantic segmentation, including attention U-Net[17], DeepLabV3+ (ref. 18), DANet[19] and MedT[20]. We applied the Dice coefficient to evaluate segmentation performance at the patch level and the agreement of predominant patterns to assess prediction at the WSI level. Comparisons were conducted with fivefold cross-validation for the TRACERx 100 cohort ($n = 53$) and on a subset of the LATTICe-A cohort ($n = 50$), an independent dataset to the training dataset.

Growth pattern and grading inference

Each WSI was divided into tiles of 2,000 × 2,000 pixels with the magnification downsampled to ×20 (approximately 0.45 µm per pixel)[29]. Each tile was then normalized to a target image to align the color before feeding it to the well-trained deep learning model, which, in turn, generated corresponding masks for all growth pattern regions detected at the pixel level. The tile masks were then stitched and further downsampled to ×1.25 (approximately 7.2 µm per pixel). Small components were empirically removed as postprocessing; lepidic patterns that were less than approximately 0.05 mm^2, and papillary, cribriform and solid patterns that were less than approximately 0.015 mm^2 were removed.

The predominant pattern and grading were inferred from a stitched and downsampled mask (approximately 7.2 µm per pixel). The growth pattern proportion for each tumor was computed as the proportion across all slides of a given tumor:

$$g_j = \frac{\sum_{i=1}^{m} S_{ij}}{\sum_{i=1}^{m} \sum_{j}^{n=6} S_{ij}}$$

$$P = \mathrm{argmax}(g_j)$$

where g_j is a proportion for the j pattern, j represents lepidic, acinar, papillary, cribriform, micropapillary and solid, i is the i-th slide, m is the number of slides per tumor, n is the number of patterns and S_{ij} is the number of pixels identified for the j pattern with the i-th slide. The predominant pattern, P, is determined as the pattern with the highest proportion. The growth pattern grading driven by AI followed the IASLC grading system[2]: grade 1, lepidic-predominant tumors with less than 20% of high-grade patterns (solid, micropapillary, cribriform); grade 2, acinar-predominant or papillary-predominant tumors with less than 20% of high-grade patterns; and grade 3, any tumor with 20% or more high-grade patterns.

Agreement between AI and pathological scores with regard to predominant patterns

The strongest correlation for growth pattern proportion between the AI and manual estimates was observed for the solid pattern (TRACERx 421, rho = 0.79; LATTICe-A correlations against each pathologist's scoring, rho1 = 0.80, rho2 = 0.77, rho3 = 0.78; TCGA, rho = 0.67; Fig. 2b and Supplementary Table 1), followed by acinar (TRACERx 421, rho = 0.69; LATTICe-A, rho1 = 0.67, rho2 = 0.58, rho3 = 0.65; TCGA, rho = 0.56; Fig. 2b and Supplementary Table 1). A moderate correlation was observed for the micropapillary subtype (TRACERx 421, rho = 0.35; LATTICe-A, rho1 = 0.35, rho2 = 0.42, rho3 = 0.40; TCGA, rho = 0.44; Fig. 2b and Supplementary Table 1). Compared with other patterns, solid-predominant tumors had the highest agreement levels between AI and manual scoring (TRACERx 421, 85.5%; LATTICe-A, 85.4%, 79.9%, 85.3% against three pathologists; TCGA, 72%; DHMC, 90.2%; Fig. 2c). A lower agreement rate was observed for micropapillary-predominant tumors (TRACERx 421, 0%; LATTICe-A, 38.1%, 19.7% and 50% against three pathologists; TCGA, 20%; DHMC, 0%; Fig. 2c). Most discrepant micropapillary-predominant cases were identified as papillary and acinar by AI (TRACERx 421, 40%; LATTICe-A, 42.8%, 63.1%, 43.7%; TCGA, 60%; DHMC, 100%), suggesting that micropapillary islands frequently mixed with acinar or papillary in micropapillary-predominant tumors.

C-index measuring prognostic ability

We used the C-index to measure the prognostic ability of the survival models. Cox regression models were considered for predicting DFS; specifically, the baseline model included age, sex, tumor stage (excluded for stage I tumors as the stage information remains the same). The AI grading-based model included clinical baseline characteristics and automated IASLC grading. The manual grading-based model included clinical baseline characteristics together with a pathologist's manual IASLC grading. When excluding clinical parameters, AI grading achieved a comparable C-index with pathological grading in stage I (TRACERx 421: AI = 0.588, 95% CI = 0.483–0.692; path = 0.593, 95% CI = 0.461–0.724; LATTICe-A: AI = 0.616, 95% CI = 0.571–0.661; path 1 = 0.609, 95% CI = 0.563–0.656; path 2 = 0.593, 95% CI = 0.545–0.641; path 3 = 0.571, 95% CI = 0.483–0.658; Supplementary Table 6) and stage I–III tumors (TRACERx 421: AI = 0.588, 95% CI = 0.547–0.630; path = 0.581, 95% CI = 0.530–0.632; LATTICe-A: AI = 0.577, 95% CI = 0.554–0.600; path1 = 0.577, 95% CI = 0.552–0.603; path 2 = 0.574, 95% CI = 0.551–0.597; path 3 = 0.569, 95% CI = 0.546–0.591; Supplementary Table 6).

Acinar morphological features

The pixel number and solidity index, that is, the proportion of pixels in the convex hull that were also in a region of interest, were used to measure the individual acinar island area and shape generated by the AI method. A higher solidity index indicated a more regular shape. The average area and solidity index of all the individual acinar islands identified from the available slides were taken as the tumor-level features.

Acinar scattering score

We adapted an established score, standard distance[32], to measure the spatial distribution of acinar patterns, which we termed 'acinar scattering':

$$d = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_0)^2 + \sum_{i=1}^{n}(y_i - y_0)^2}{n \times N}}$$

where d is the standard distance, n is the number of isolated acinar islands within the tissue identified by the proposed AI method, N is the area of the tissue, (x_i, y_i) is the centroid of an acinar island and (x_0, y_0) is the mean center of all the acinar islands.

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{n}, \, y_0 = \frac{\sum_{i=1}^{n} y_i}{n}$$

A higher acinar scattering score indicated a more scattered distribution of acini across the tissue. The median value of all available slides for a given tumor was taken as the tumor-level score. The optimal cutoff (0.36) separating tumors into low-scattering and high-scattering groups was selected from the discovery cohort, LATTICe-A, which was then applied directly to the TRACERx 421 cohort.

In a univariable model, acinar scattering was prognostic of DFS for LATTICe-A in grade 2 and 3 tumors, respectively (grade 2, $n = 212$, $P = 1.95 \times 10^{-5}$, HR = 2.48, 95% CI = 1.63–3.76; grade 3, $n = 570$, $P = 0.007$, HR = 1.35, 95% CI = 1.08–1.68; Extended Data Fig. 8b,c), but not in grade 1 tumors (Extended Data Fig. 8a). In the TRACERx 421 cohort, high acinar scattering was associated with reduced DFS in grade 3 tumors ($n = 137$, $P = 0.042$, HR = 1.64, 95% CI = 1.01–2.65) and remained borderline in grade 2 tumors ($n = 56$, $P = 0.053$, HR = 2.74, 95% CI = 0.99–7.61), but was not notable in grade 1 tumors. The lack of statistical significance was probably due to the smaller number of patients and events in the grade 1 subgroup. When merging grade 1 and 2 tumors, the prognostic effect of acinar scattering was observed (TRACERx 421, $n = 68$, $P = 0.025$, HR = 2.79, 95% CI = 1.14–6.85; LATTICe-A, $n = 267$, $P = 1.39 \times 10^{-5}$, HR = 2.36, 95% CI = 1.60–3.48; Extended Data Fig. 8d).

Statistics and reproducibility

Correlation tests used Spearman's method and were generated using the function cor.test from the stats v.4.1.2R package. Confusion matrices were obtained using the function confusionMatrix from the caret v.6.0-93R package. Fleiss' kappa was computed to assess the agreement among observers using the function kappam.fleiss from the irr v.0.84.1R package. Survival analyses were conducted using the Kaplan–Meier estimator (ggsurvplot R function from the survminer v.0.4.9 and survival v.3.2-13R packages) as well as the Cox model (coxph R function, displayed using the ggforest R function). The differences between grade strata Kaplan–Meier curves were determined using Wald tests. Forest plots showed the HR on the *x* axis; each variable's HR was plotted and annotated with a 95% CI. All HRs were computed for all time points (the whole survival curve was not at a specific time point). For statistical comparisons among groups, a two-sided, nonparametric, unpaired Wilcoxon rank-sum test was used for the continuous variables, while a Fisher's exact test was used for the categorical variables. A Kruskal–Wallis test was used for comparisons among over two groups, unless stated otherwise. Predictive performance was assessed using a C-index[33] within 5 years, computed with the function Inf.Cval from the survC1 v.1.0-3R package. Multicollinearity between AI and manual grading, and between two manual gradings were assessed using the function vif from the car v.3.0-12R package. All statistical tests were two-sided and $P < 0.05$ was considered as statistically significant. To adjust P values for multiple comparisons, the Benjamini–Hochberg method was used. The packages tidyverse v.2.0.0 and tidyr v.1.3.0 were used for data processing in R. Plotting was done using ggplot2 v.3.4.1, RColorBrewer v.1.1-3 and ggpubr v.0.5.0R packages. All statistical analyses were conducted in R v.4.1.2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The training dataset consisting of annotations on small image tiles have been deposited in Zenodo (https://doi.org/10.5281/zenodo.10016027). Previously published image data that were reanalyzed in this study can be requested from https://bmirds.github.io/LungCancer/. The human LUAD diagnostic slide images were derived from the TCGA Research Network at https://portal.gdc.cancer.gov/. Images generated by the AI model in Fig. 2a and Extended Data Figs. 2, 3a and 7f can be accessed at figshare (https://doi.org/10.6084/m9.figshare.24599796). For the TRACERx study, all of the scanned diagnostic histological images have a study number label embedded in the file that prevents complete anonymization. Therefore, these images cannot be shared, in line with the ethical approval for the study. Requests for access to the TRACERx dataset for academic noncommercial research purposes can be submitted through the Cancer Research UK and UCL Cancer Trials Centre (ctc.tracerx@ucl.ac.uk) and are subject to review of a project proposal that will be evaluated by a TRACERx data access committee, entering into an appropriate data access agreement and any applicable ethical approvals. The time frame of response to requests is about 6 months. LATTICe-A study data and materials are currently subject to a material and data transfer agreement between the University of Leicester, the University of Cambridge and NHS Greater Glasgow and Clyde, which includes a restricted access period of 5 years, precluding any access by other third parties during this time. After the 5-year period, restricted access data can be accessed by application to NHS Greater Glasgow and Clyde Biorepository (clare.orange@ggc.scot.nhs.uk; john.lequesne@glasgow.ac.uk) as custodians; the data access request will be reviewed and released under their research ethics committee-approved tissue bank protocols. Requests will be reviewed and approved within 6–8 weeks and will be accompanied by a data sharing agreement detailing the conditions and restrictions of use and publication. Source data are provided with this paper.

Code availability

The AI pipeline for growth pattern segmentation is available at https://github.com/xi11/AIgrading. All code used for the analyses was developed in R v.4.1.2 and is available to reproduce all figures (https://github.com/xi11/AIgrading).

References

1. Nicholson, A. G. et al. The 2021 WHO Classification of Lung Tumors: impact of advances since 2015. *J. Thorac. Oncol.* **17**, 362–387 (2022).
2. Moreira, A. L. et al. A grading system for invasive pulmonary adenocarcinoma: a proposal from the International Association for the Study of Lung Cancer Pathology Committee. *J. Thorac. Oncol.* **15**, 1599–1610 (2020).
3. Boland, J. M., Wampfler, J. A., Yang, P. & Yi, E. S. Growth pattern-based grading of pulmonary adenocarcinoma—analysis of 534 cases with comparison between observers and survival analysis. *Lung Cancer* **109**, 14–20 (2017).
4. Thunnissen, E. et al. Reproducibility of histopathological subtypes and invasion in pulmonary adenocarcinoma. An international interobserver study. *Mod. Pathol.* **25**, 1574–1583 (2012).
5. Deng, C. et al. Validation of the novel International Association for the Study of Lung Cancer grading system for invasive pulmonary adenocarcinoma and association with common driver mutations. *J. Thorac. Oncol.* **16**, 1684–1693 (2021).
6. Tavernari, D. et al. Nongenetic evolution drives lung adenocarcinoma spatial heterogeneity and progression. *Cancer Discov.* **11**, 1490–1507 (2021).
7. Deshmukh, G. et al. FEEDNet: a feature enhanced encoder–decoder LSTM network for nuclei instance segmentation for histopathological diagnosis. *Phys. Med. Biol.* 67, https://doi.org/10.1088/1361-6560/ac8594 (2022)
8. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
9. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J. & Ciompi, F. HookNet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med. Image Anal.* **68**, 101890 (2021).
10. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
11. Wang, Y. et al. Improved breast cancer histological grading using deep learning. *Ann. Oncol.* **33**, 89–98 (2022).
12. Gertych, A. et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci. Rep.* **9**, 1483 (2019).
13. Wei, J. W. et al. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci. Rep.* **9**, 3358 (2019).
14. Pan, X. et al. in *Computational Mathematics Modeling in Cancer Analysis* 78–90 (Springer, 2022).
15. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2881–2890 (IEEE, 2017).
16. Lin, G., Milan, A., Shen, C. & Reid, I. RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1925–1934 (IEEE, 2017).
17. Oktay, O. et al. Attention U-Net: learning where to look for the pancreas. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1804.03999 (2018).
18. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder–decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018* (eds. Ferrari, V. et al.) 833–851 (2018).
19. Fu, J. et al. Dual attention network for scene segmentation. In *Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3146–3154 (IEEE, 2019).
20. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. & Patel, V. M. Medical Transformer: gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021* (eds. de Bruijne, M. et al.) 36–46 (2021).
21. Yoshizawa, A. et al. Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. *Mod. Pathol.* **24**, 653–664 (2011).
22. Karasaki, T. et al. Evolutionary characterization of lung adenocarcinoma morphology in TRACERx. *Nat. Med.* **29**, 833–845 (2023).
23. Moore, D. A. et al. In situ growth in early lung adenocarcinoma may represent precursor growth or invasive clone outgrowth—a clinically relevant distinction. *Mod. Pathol.* **32**, 1095–1105 (2019).
24. Thunnissen, E. et al. Elastin in pulmonary pathology: relevance in tumours with a lepidic or papillary appearance. A comprehensive understanding from a morphological viewpoint. *Histopathology* **80**, 457–467 (2022).
25. Thunnissen, E. et al. Defining morphologic features of invasion in pulmonary nonmucinous adenocarcinoma with lepidic growth: a proposal by the International Association for the Study of Lung Cancer Pathology Committee. *J. Thorac. Oncol.* **18**, 447–462 (2023).
26. Frankell, A. M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* **616**, 525–533 (2023).
27. *WHO Classification of Tumours: Thoracic Tumours* (WHO, 2021).
28. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
29. AbdulJabbar, K. et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
30. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
31. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
32. Mitchell, A. & Griffin, L. S. *The ESRI Guide to GIS Analysis, Vol. 2. Spatial Measurements and Statistics* (ESRI, 2020).
33. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).

Acknowledgements

This study was funded by a Cancer Research UK Career Establishment Award to Y.Y. (C45982/A21808). The TRACERx study (ClinicalTrials.gov registration: NCT01888601) is sponsored by University College London (UCL) (UCL/12/0279) and has been approved by an independent research ethics committee (13/LO/1546). TRACERx is funded by Cancer Research UK (C11496/A17786) and coordinated through the Cancer Research UK and UCL Cancer Trials Centre. Y.Y. acknowledges additional support from Breast Cancer Now (2015NovPR638), the Children's Cancer and Leukaemia Group (CCLGA201906), National Institutes of Health (NIH) grant nos. U54 CA217376 and R01 CA185138, Congressionally Directed Medical Research Programs, Breast Cancer Research Program Award no. BC132057, European Commission Innovative Training Network (H2020-MSCA-ITN-2019), Wellcome Trust (105104/Z/14/Z) and the Royal Marsden/Institute of Cancer Research National Institute of Health Research Biomedical Research Centre. C.S. is a Royal Society Napier Research Professor (RSRP\R\210001). This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2041), the UK Medical Research Council (MRC) (CC2041) and the Wellcome Trust (CC2041). For the purpose of open access, the authors have applied a CC BY public copyright license to any author's accepted manuscript version arising from this submission. C.S. is funded by Cancer Research UK (TRACERx, C11496/A17786), PEACE (C416/A21999) and Cancer Research UK Cancer Immunotherapy Catalyst Network; Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust; the Butterfield and Stoneygate Trust; the NovoNordisk Foundation (ID16584); a Royal Society Professorship Enhancement Award (RP/EA/180007); the National Institute for Health Research University College London Hospitals Biomedical Research Centre; the Cancer Research UK-University College London Centre; the Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (BCRF-22-157); a Cancer Research UK Early Detection an Diagnosis Primer Award (grant no. EDDPMA-Nov21/100034); and a Mark Foundation for Cancer Research Aspire Award (grant no. 21-029-ASP). Research grants are administered by the American Association for Cancer Research, the scientific partner of SU2C. C.S. is in receipt of an European Research Council (ERC) Advanced Grant (PROTEUS) under the European Union's Horizon 2020 research and innovation program (grant no. 835297). We thank L. Officer-Jones, C. Ficken and M. Das from the MRC Toxicology Unit Core Histology facility for their expert technical assistance. Figure 1a and Extended Data Figs. 1c and 3b,c were first published in Pan et al.[14].

Author contributions

X.P. collated the pathological annotations and the data from the cohorts, developed the AI method and performed the analyses. K.A. collated the pathological annotations and the data from the LATTICe-A cohort. J.C.-L. reviewed the TCGA cohort and helped write the manuscript. A.-I.G. collated the pathological annotations and helped write the manuscript. H.Z. conducted the comparison of the different AI methods. D.A.M., A.H.K.C. and J.L.Q. provided the pathological annotations. J.L.Q., D.A.M. and J.B. reviewed the LATTICe-A cohort. J.L.Q., M.S. and C.R.W. provided data and advice for the LATTICe-A cohort. T.K. collated the clinicopathological data of the TRACERx 421 cohort and helped analyze the data. S.V. performed the histology sample generation and digitized the H&E slides for the TRACERx 421 cohort. A.H. performed the survival analyses for the TRACERx 421 cohort and helped analyze the data. C.S. and M.J.-H. provided clinical expertise and oversight of the TRACERx study. J.L.Q., D.A.M. S.J.A. and A.G.N. provided histopathological expertise and helped write the manuscript. D.A.M. led the central pathology review and collated the pathology data for the TRACERx 421 cohort. J.L.Q. led the central pathology review and collated the pathology data for the LATTICe-A cohort. Y.Y., J.L.Q. and D.A.M. jointly conceived and supervised the study. X.P., K.A. and Y.Y. wrote the manuscript with input from all authors.

Competing interests

S.V. is a coinventor to a patent of methods for detecting molecules in a sample (patent no. 10578620). A.H. has received fees from Abbvie, Almirall, Boehringer Ingelheim, Clovis Oncology, Ipsen, Takeda Pharmaceuticals, AstraZeneca, Daiichi Sankyo, Merck Serono, Merck/MSD, UCB, Kyowa Kirin, Servier, Sobi, Pfizer and Roche for delivering general education and training in clinical trials; has received fees for member of independent data monitoring committees for Roche-sponsored clinical trials and academic projects on real-world evidence or tumor-agnostic therapies coordinated by Roche; he has been paid honoraria for speaking at Roche-funded conferences (on real-world data); he has an academic collaboration with Navio and is an unpaid member of their advisory board; he is an investigator for an academic study (SUMMIT) sponsored by UCL, which is funded by GRAIL; he has received one honorarium for an advisory board meeting for GRAIL; he has received a consulting fee from Evidera (for one GRAIL-initiated project); and he has previously owned shares in Illumina and Thermo Fisher Scientific (sold in 2020); he is on the scientific advisory board for Adela Bio and has received no payments or honoraria for this, although he has share options available. A.G.N. reports personal fees from Merck, Boehringer Ingelheim, Novartis, AstraZeneca, Bristol Myers Squibb, Roche, Abbvie, Oncologica, Uptodate, the European Society of Oncology, Takeda Pharmaceuticals, Sanofi and Liberium, as well as personal fees and grants from Pfizer. M.J-H. is a Cancer Research UK Career Establishment Awardee and has received funding from Cancer Research UK, the International Association for the Study of Lung Cancer and International Lung Cancer Foundation, the Lung Cancer Research Foundation, the Rosetrees Trust, UK and Ireland Neuroendocrine Tumour Society, the National Institute for Health Research (NIHR) and the NIHR UCLH Biomedical Research Centre. M.J-H. has consulted for, and is a member of, the Achilles Therapeutics Scientific advisory board and steering committee, has received speaker honoraria from Pfizer, Astex Pharmaceuticals and Oslo Cancer Cluster, and holds a patent (no. PCT/US2017/028013) relating to methods for lung cancer detection. C.S. acknowledges grant support from AstraZeneca, Boehringer Ingelheim, Bristol Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx, collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical. He is an AstraZeneca advisory board member and chief investigator for the AZ MeRmaiD 1 and 2 clinical trials; he is also co-chief investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's scientific advisory board. He receives consultant fees from Achilles Therapeutics (scientific advisory board member), Bicycle Therapeutics (scientific advisory board), Genentech, Medicxi, Roche Innovation Centre-Shanghai, Metabomed (until July 2022) and the Sarah Cannon Research Institute. C.S. has received honoraria from Amgen, AstraZeneca, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, Illumina and Roche-Ventana. C.S. had stock options in Apogen Biotechnologies and GRAIL until June 2021; he currently has stock options in Epic Bioscience, Bicycle Therapeutics; he has stock options and is a cofounder of Achilles Therapeutics. C.S. holds patents relating to assay technology to detect tumor recurrence (no. PCT/GB2017/053289), target neoantigens (no. PCT/EP2016/059401), identify patent response to immune checkpoint blockade (no. PCT/EP2016/071471), determine HLA loss of heterozygosity (no. PCT/GB2018/052004), predict survival rates of patients with cancer (no. PCT/GB2020/050221) and identify patients who respond to cancer treatment (no. PCT/GB2018/051912), as well as a US patent related to detecting tumor mutations (no. PCT/US2017/28013) and methods for lung cancer detection (no. US20190106751A1), and both European and US patents related to identifying insertion and deletion mutation targets (no. PCT/GB2018/051892). Y.Y. has received speaker's bureau honoraria from Roche and consulted for Merck. D.A.M. reports speaker fees from Eli Lilly, AstraZeneca and Takeda Pharmaceuticals, consultancy fees from AstraZeneca, Thermo Fisher Scientific, Takeda Pharmaceuticals, Amgen, Janssen, MIM Software, Bristol Myers Squibb and Eli Lilly, and has received educational support from Takeda Pharmaceuticals and Amgen. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-023-00694-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-023-00694-w.

Correspondence and requests for materials should be addressed to Yinyin Yuan, John Le Quesne or David A. Moore.

Peer review information *Nature Cancer* thanks Lynette Sholl and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

[1]Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK. [2]Division of Molecular Pathology, The Institute of Cancer Research, London, UK. [3]Medical Research Council Toxicology Unit, University of Cambridge, Cambridge, UK. [4]Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [5]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK. [6]Leicester Cancer Research Centre, University of Leicester, Leicester, UK. [7]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [8]Hope Against Cancer and Leicester Experimental Cancer Medicine Centre, Leicester, UK. [9]Institute for Lung Health, NIHR Leicester Biomedical Research Centre, Leicester, UK. [10]Cancer Research UK & UCL Cancer Trials Centre, London, UK. [11]Department of Histopathology, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [12]National Heart and Lung Institute, Imperial College London, London, UK. [13]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK. [14]Department of Medical Oncology, University College London Hospitals NHS Foundation Trust, London, UK. [15]Molecular Pathology, School of Cancer Sciences, University of Glasgow, Glasgow, UK. [16]Cancer Research UK Beatson Institute of Cancer Research, Glasgow, UK. [17]NHS Greater Glasgow and Clyde, Glasgow, UK. [18]Department of Cellular Pathology, University College London Hospitals, London, UK. [19]Present address: Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [20]Present address: AstraZeneca Computational Pathology, Munich, Germany. [108]These authors contributed equally: Khalid AbdulJabbar, Jose Coelho-Lima. [109]These authors jointly supervised this work: Yinyin Yuan, John Le Quesne, David A. Moore. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: yyuan6@mdanderson.org; john.lequesne@glasgow.ac.uk; d.moore@ucl.ac.uk

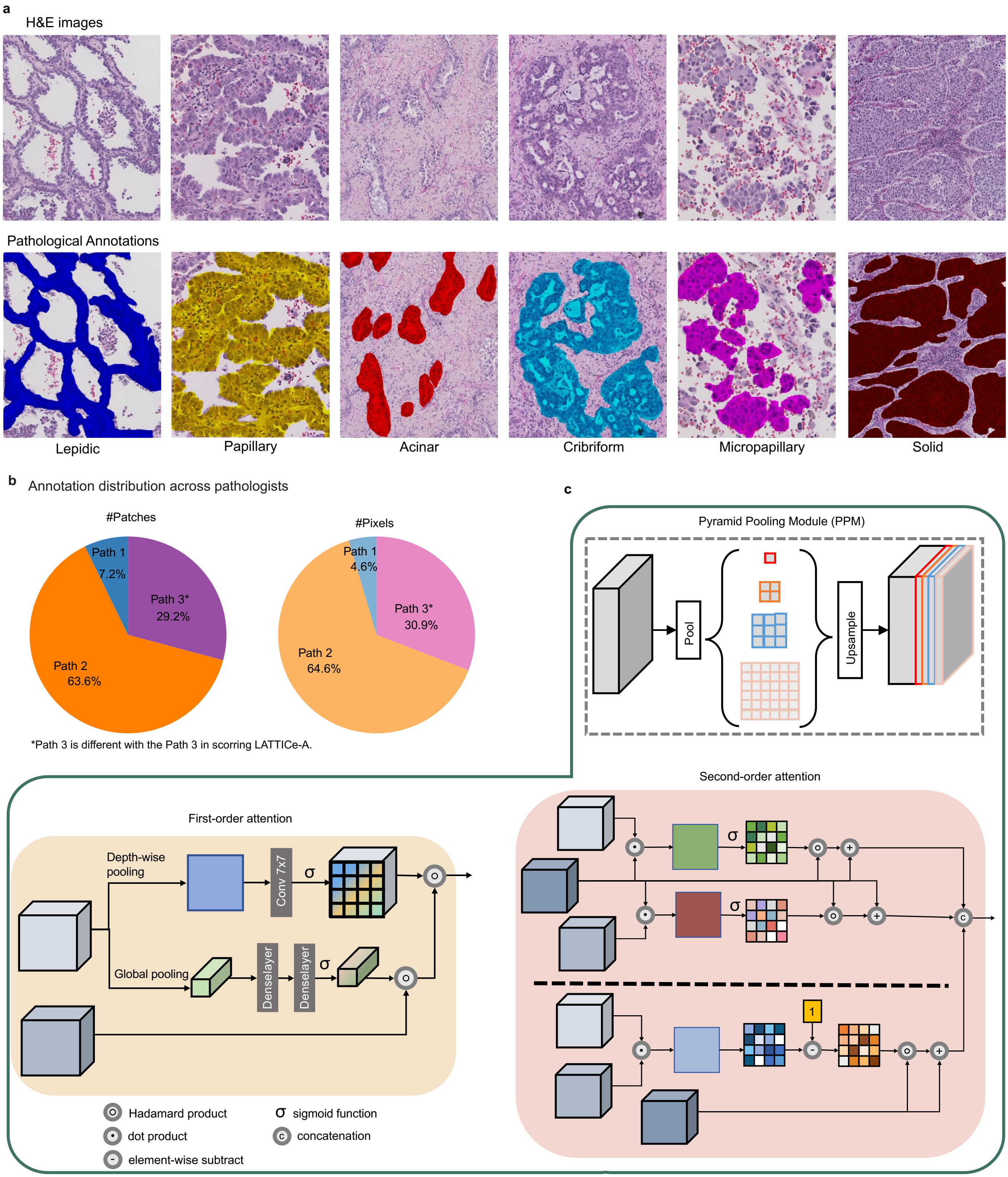
TRACERx Consortium

Charles Swanton[5,7,14], Mariam Jamal-Hanjani[7,13,14], Hanyun Zhang[1,2], Khalid AbdulJabbar[1,2,108], Xiaoxi Pan[1,2,19], Yinyin Yuan[1,2,19,109], Allan Hackshaw[10], John Le Quesne[15,16,17], Selvaraju Veeriah[5,7], Takahiro Karasaki[5,7], Sam M. Janes[21], Anne-Marie Hacker[10], Abigail Sharp[10], Sean Smith[10], Harjot Kaur Dhanda[10], Kitty Chan[10], Camilla Pilotti[10], Rachel Leslie[10], Anca-Ioana Grapa[1,2], David Chuter[22], Mairead MacKenzie[22], Serena Chee[23], Aiman Alzetani[23], Eric Lim[24,25], Paulo De Sousa[25], Simon Jordan[25], Alexandra Rice[25], Hilgardt Raubenheimer[25], Harshil Bhayani[25], Lyn Ambrose[25], Anand Devaraj[25], Hema Chavan[25], Sofina Begum[25], Silviu I. Buderi[25], Daniel Kaniu[25], Mpho Malima[25], Sarah Booth[25], Nadia Fernandes[25], Pratibha Shah[25], Chiara Proli[25], Madeleine Hewish[26,27], Sarah Danson[28], Michael J. Shackcloth[29], Lily Robinson[30], Peter Russell[30], Kevin G. Blyth[31,32,33], Andrew Kidd[34], Alan Kirk[35], Mo Asif[35], Rocco Bilancia[35], Nikos Kostoulas[35], Mathew Thomas[35], Andrew G. Nicholson[11,12], Craig Dick[17], Jason F. Lester[36], Amrita Bajaj[37], Apostolos Nakas[37], Azmina Sodha-Ramdeen[37], Mohamad Tufail[37], Molly Scotland[37], Rebecca Boyles[37], Sridhar Rathinam[37], Dean A. Fennell[37,38], Claire Wilson[6], Domenic Marrone[38], Sean Dulloo[38], Gurdeep Matharu[39], Jacqui A. Shaw[39], Joan Riley[39], Lindsay Primrose[39], Ekaterini Boleti[40], Heather Cheyne[41], Mohammed Khalil[41], Shirley Richardson[41], Tracey Cruickshank[41], Gillian Price[42,43], Keith M. Kerr[43,44], Sarah Benafif[45], Dionysis Papadatos-Pastos[45], James Wilson[45], Tanya Ahmad[45], Jack French[46], Kayleigh Gilbert[46], Babu Naidu[47], Akshay J. Patel[48], Aya Osman[48], Christer Lacson[48], Gerald Langman[48], Helen Shackleford[48], Madava Djearaman[48], Salma Kadiri[49], Gary Middleton[48,50], Angela Leek[51], Jack Davies Hodgkinson[51], Nicola Totten[51], Angeles Montero[52], Elaine Smith[52], Eustace Fontaine[52], Felice Granato[52], Juliette Novasio[52], Kendadai Rammohan[52], Leena Joseph[52], Paul Bishop[52], Rajesh Shah[52], Stuart Moss[52], Vijay Joshi[52], Philip Crosbie[52,53,54], Antonio Paiva-Correia[55], Anshuman Chaturvedi[54,56], Lynsey Priest[54,56], Pedro Oliveira[54,56], Fabio Gomes[56], Kate Brown[56], Mathew Carter[56], Colin R. Lindsay[57], Fiona H. Blackhall[57], Matthew G. Krebs[57], Yvonne Summers[57], Alexandra Clipson[54,58], Jonathan Tugwood[54,58], Alastair Kerr[54,58], Dominic G. Rothwell[54,58], Caroline Dive[54,58], Hugo J. W. L. Aerts[49,59,60], Roland F. Schwarz[61,62], Tom L. Kaufmann[62,63], Peter Van Loo[64,65,66], Gareth A. Wilson[5], Rachel Rosenthal[5], Andrew Rowan[5], Chris Bailey[5], Claudia Lee[5], Emma Colliver[5], Katey S. S. Enfield[5], Mark S. Hill[5], Mihaela Angelova[5], Oriol Pich[5], Michelle Leung[5,7,67], Alexander M. Frankell[5,7], Crispin T. Hiley[5,7], Emilia L. Lim[5,7], Haoran Zhai[5,7], Maise Al Bakir[5,7], Nicolai J. Birkbak[5,7,68,69,70], Olivia Lucas[5,7,71,72], Ariana Huebner[5,7,67], Clare Puttick[5,7,67], Kristiana Grigoriadis[5,7,67], Michelle Dietzen[5,7,67], David A. Moore[5,7,18,109], Dhruva Biswas[5,7,73], Foteini Athanasopoulou[5,7,74], Sophia Ward[5,7,74], Jonas Demeulemeester[66,75,76], Carla Castignani[66,77], Elizabeth Larose Cadieux[66,77], Judit Kisistok[68,69,70], Mateo Sokac[68,69,70], Zoltan Szallasi[78,79,80], Miklos Diossy[78,79,81], Roberto Salgado[82,83], Aengus Stewart[84], Alastair Magness[84], Clare E. Weeden[84], Dina Levi[84], Eva Grönroos[84], Imran Noorani[84], Jacki Goldman[84], Mickael Escudero[84], Philip Hobson[84], Roberto Vendramin[84], Stefan Boeing[84],

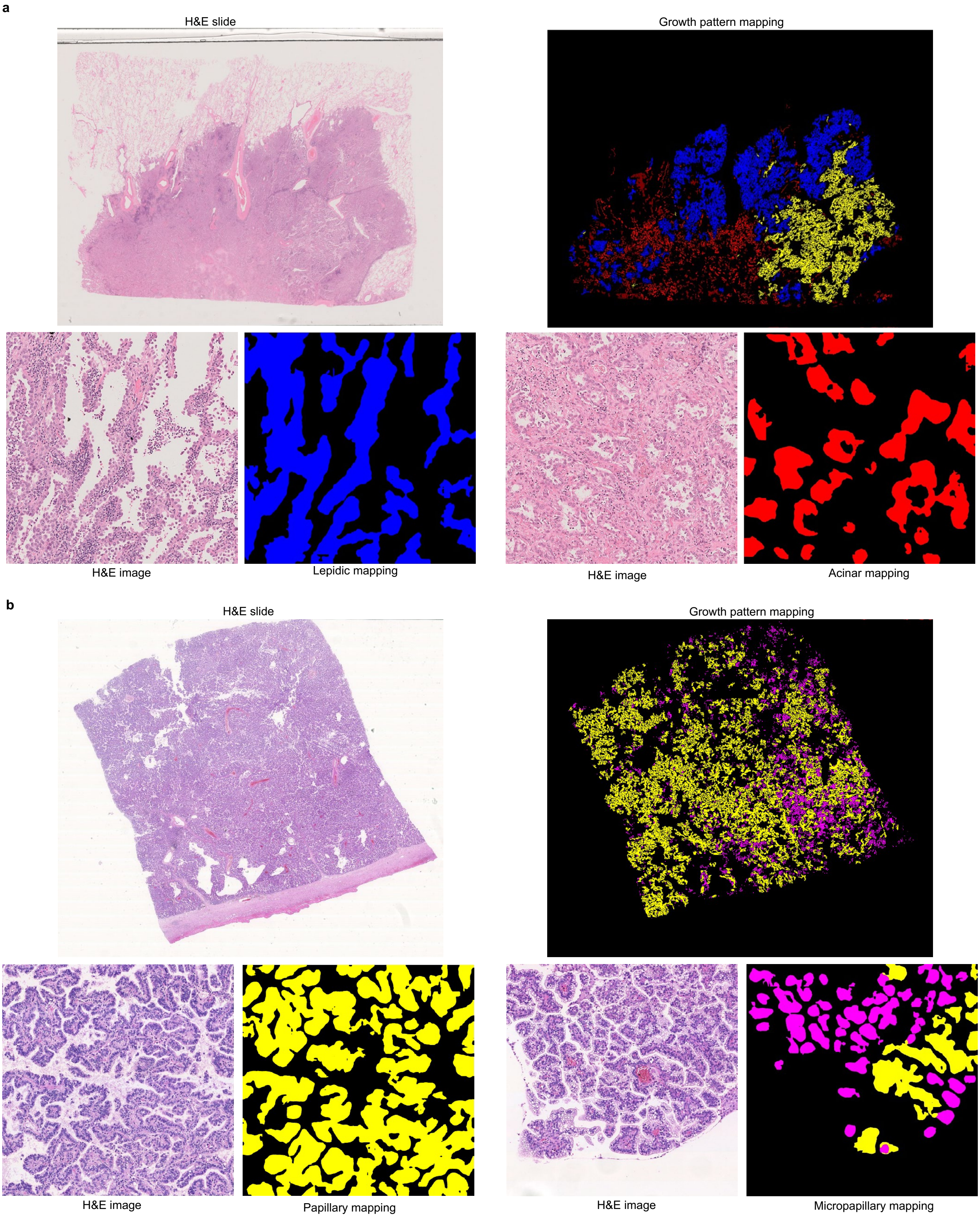
Tamara Denner[84], Vittorio Barbè[84], Wei-Ting Lu[84], William Hill[84], Yutaka Naito[84], Zoe Ramsden[84], George Kassiotis[84,85], Angela Dwornik[86], Angeliki Karamani[86], Benny Chain[86], David R. Pearce[86], Despoina Karagianni[86], Felip Gálvez-Cancino[86], Georgia Stavrou[86], Gerasimos Mastrokalos[86], Helen L. Lowe[86], Ignacio Garcia Matos[86], James L. Reading[86], John A. Hartley[86], Kayalvizhi Selvaraju[86], Kezhong Chen[86], Leah Ensell[86], Mansi Shah[86], Maria Litovchenko[86], Olga Chervova[86], Piotr Pawlik[86], Robert E. Hynds[86], Samuel Gamble[86], Seng Kuong Anakin Ung[86], Supreet Kaur Bola[86], Victoria Spanswick[86], Yin Wu[86], Othman Al-Sawaf[86,87], Thomas Patrick Jones[67], Stephan Beck[77], Miljana Tanic[77,88], Teresa Marafioti[18], Elaine Borg[18], Mary Falzon[18], Reena Khiroya[18], Antonia Toncheva[7], Christopher Abbosh[7], Corentin Richard[7], Cristina Naceur-Lombardelli[7], Francisco Gimeno-Valiente[7], Krupa Thakkar[7], Mariana Werner Sunderland[7], Monica Sivakumar[7], Nnennaya Kanu[7], Paulina Prymas[7], Sadegh Saghafinia[7], Sharon Vanloo[7], Jie Min Lam[7,13,45], Wing Kin Liu[7,13], Abigail Bunkum[7,13,71], Sonya Hessey[7,13,71], Simone Zaccaria[7,71], Carlos Martínez-Ruiz[7,67], James R. M. Black[7,67], Kerstin Thol[7,67], Robert Bentham[7,67], Kevin Litchfield[7,89], Nicholas McGranahan[7,67], Sergio A. Quezada[7,90], Martin D. Forster[7,45], Siow Ming Lee[7,45], Javier Herrero[73], Emma Nye[91], Richard Kevin Stone[91], Jerome Nicod[74], Jayant K. Rane[5,86], Karl S. Peggs[92,93], Kevin W. Ng[94], Krijn Dijkstra[95,96], Matthew R. Huska[97], Emilie Martinoni Hoogenboom[72], Fleur Monk[72], James W. Holding[72], Junaid Choudhary[72], Kunal Bhakhri[72], Marco Scarci[72], Pat Gorman[72], Robert C. M. Stephens[72], Yien Ning Sophia Wong[72], Zoltan Kaplar[72], Steve Bandula[72], Thomas B. K. Watkins[5,86], Catarina Veiga[98], Gary Royle[99], Charles-Antoine Collins-Fekete[100], Francesco Fraioli[101], Paul Ashford[102], Alexander James Procter[103], Asia Ahmed[103], Magali N. Taylor[103], Arjun Nair[103,104], David Lawrence[105], Davide Patrini[105], Neal Navani[106,107] & Ricky M. Thakrar[106,107]

[21]Lungs for Living Research Centre, UCL Respiratory, Department of Medicine, University College London, London, UK. [22]Independent Cancer Patient's Voice, London, UK. [23]University Hospital Southampton NHS Foundation Trust, Southampton, UK. [24]Academic Division of Thoracic Surgery, Imperial College London, London, UK. [25]Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [26]Royal Surrey Hospital, Royal Surrey Hospitals NHS Foundation Trust, Guildford, UK. [27]University of Surrey, Guildford, UK. [28]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [29]Liverpool Heart and Chest Hospital, Liverpool, UK. [30]Princess Alexandra Hospital, The Princess Alexandra Hospital NHS Trust, Harlow, UK. [31]School of Cancer Sciences, University of Glasgow, Glasgow, UK. [32]Beatson Institute for Cancer Research, University of Glasgow, Glasgow, UK. [33]Queen Elizabeth University Hospital, Glasgow, UK. [34]Institute of Infection, Immunity & Inflammation, University of Glasgow, Glasgow, UK. [35]Golden Jubilee National Hospital, Clydebank, UK. [36]Singleton Hospital, Swansea Bay University Health Board, Swansea, UK. [37]University Hospitals of Leicester NHS Trust, Leicester, UK. [38]University of Leicester, Leicester, UK. [39]Cancer Research Centre, University of Leicester, Leicester, UK. [40]Royal Free London NHS Foundation Trust, London, UK. [41]Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [42]Department of Medical Oncology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [43]University of Aberdeen, Aberdeen, UK. [44]Department of Pathology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [45]Department of Oncology, University College London Hospitals, London, UK. [46]The Whittington Hospital NHS Trust, London, UK. [47]Birmingham Acute Care Research Group, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. [48]University Hospital Birmingham NHS Foundation Trust, Birmingham, UK. [49]Artificial Intelligence in Medicine AIM Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. [50]Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. [51]Manchester Cancer Research Centre Biobank, Manchester, UK. [52]Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, UK. [53]Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK. [54]Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK. [55]Manchester University NHS Foundation Trust, Manchester, UK. [56]The Christie NHS Foundation Trust, Manchester, UK. [57]Division of Cancer Sciences, The University of Manchester and The Christie NHS Foundation Trust, Manchester, UK. [58]Cancer Research UK Manchester Institute Cancer Biomarker Centre, University of Manchester, Manchester, UK. [59]Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [60]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands. [61]Institute for Computational Cancer Biology, Center for Integrated Oncology, Cancer Research Center Cologne Essen, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [62]Berlin Institute for the Foundations of Learning and Data, Berlin, Germany. [63]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. [64]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [65]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [66]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. [67]Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [68]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [69]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [70]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [71]Computational Cancer Genomics Research Group, University College London Cancer Institute, London, UK. [72]University College London Hospitals, London, UK. [73]Bill Lyons Informatics Centre, University College London Cancer Institute, London, UK. [74]Advanced Sequencing Facility, The Francis Crick Institute, London, UK. [75]Integrative Cancer Genomics Laboratory, Department of Oncology, KU Leuven, Leuven, Belgium. [76]VIB-KU Leuven Center for Cancer Biology, Leuven, Belgium. [77]Medical Genomics, University College London Cancer Institute, London, UK. [78]Danish Cancer Society Research Center, Copenhagen, Denmark. [79]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [80]Department of Bioinformatics, Semmelweis University, Budapest, Hungary. [81]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. [82]Department of Pathology, Ziekenhuis aan de Stroom Hospitals, Antwerp, Belgium. [83]Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [84]The Francis Crick Institute, London, UK. [85]Department of Infectious Disease, Faculty of Medicine, Imperial College London, London, UK. [86]University College London Cancer Institute, London, UK. [87]Department I of Internal Medicine, University Hospital of Cologne, Cologne, Germany. [88]Experimental Oncology, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia. [89]Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK. [90]Immune Regulation and Tumour Immunotherapy Group, Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [91]Experimental Histopathology, The Francis Crick Institute, London, UK. [92]Department of Haematology, University College London Hospitals, London, UK. [93]Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [94]Retroviral Immunology Group, The Francis Crick Institute, London, UK. [95]Department of Molecular Oncology and Immunology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [96]Oncode Institute, Utrecht, the Netherlands. [97]Bioinformatics and Systems Biology, Method Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany. [98]Department of Medical Physics and Biomedical Engineering, Centre for Medical Image Computing, London, UK. [99]Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, UK. [100]Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [101]Institute of Nuclear Medicine, Division of Medicine, University

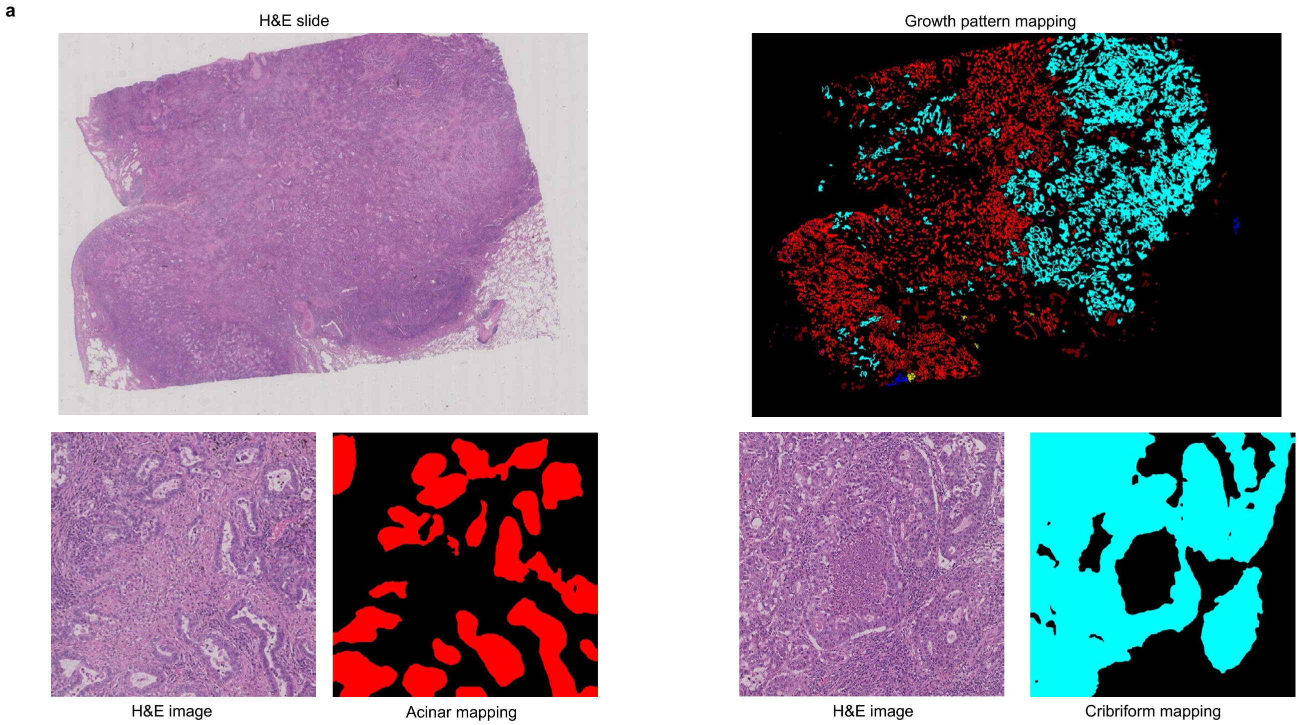
College London, London, UK. [102]Institute of Structural and Molecular Biology, University College London, London, UK. [103]Department of Radiology, University College London Hospitals, London, UK. [104]University College London Respiratory, Department of Medicine, University College London, London, UK. [105]Department of Thoracic Surgery, University College London Hospital NHS Trust, London, UK. [106]Lungs for Living Research Centre, University College London Respiratory, University College London, London, UK. [107]Department of Thoracic Medicine, University College London Hospitals, London, UK.



Extended Data Fig. 1 | Precise pathological annotations for training and sub-modules of the developed deep learning model (ANORAK). a. Examples illustrating morphologically distinct growth patterns in lung adenocarcinoma. **b.** Distribution of annotations regarding the number of patches and pixels. **c.** Detailed architectures of sub-modules developed for the AI method.



Extended Data Fig. 2 | Segmentation performance. a,**b**. Segmentations generated by AI at low-power and high-power resolutions, deposited in 10.6084/m9.figshare.24599796.



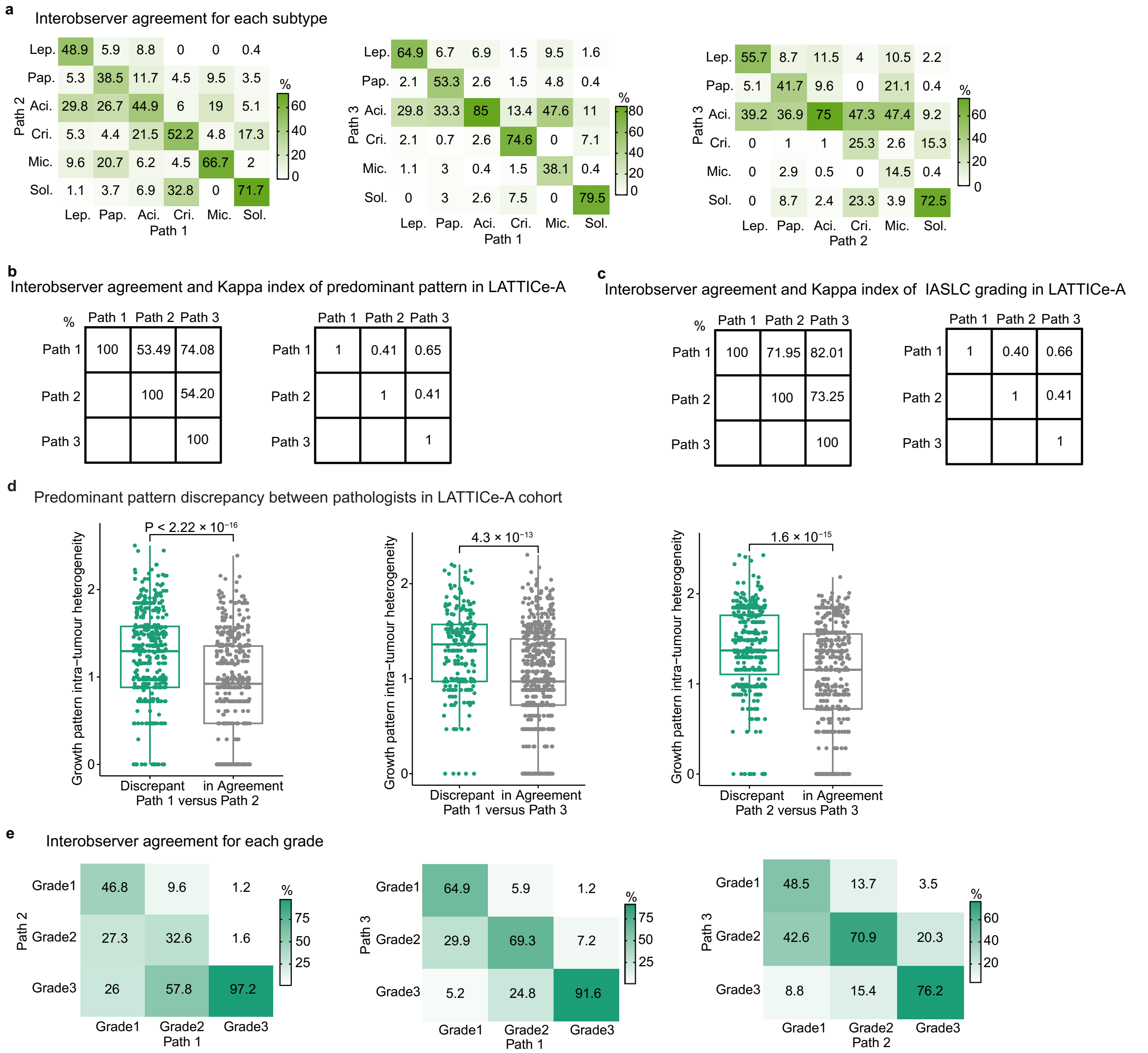
b Performance comparison with different integration methods at patch level (Dice)

Method	Lepidic	Papillary	Acinar	Cribriform	Micropapillary	Solid	Avgerage
Single Stream	0.5539	0.4755	0.5686	0.2921	0.4449	0.6637	0.4998
Multi-ADD	0.5543	0.5062	0.6104	0.3001	0.4318	0.6793	0.5137
Multi-FO	0.6439	0.6101	0.6074	0.3150	0.4117	0.6647	0.5421
Multi-SO	0.6896	0.5419	0.6253	0.4732	0.4733	0.7278	0.5885
Mutli-FO&SO (ANORAK)	0.7463	0.5863	0.6310	0.4966	0.4430	0.7170	0.6034

c Performance comparison with different segmentation methods at patch (Dice) and WSI (agreement) levels

Method	Lepidic	Papillary	Acinar	Cribriform	Micropapillary	Solid	Patch-level Dice TRACERx 100	WSI-level Agreement (%)		#Parameters
								TRACERx 100, n=53	LATTICe-A, n = 50	
attention U-Net	0.4806	0.4269	0.4518	0.1899	0	0.7126	0.3770	48.98	42.00	15.55M
DeepLabV3+	0.6864	0.5186	0.6095	0.4194	0.4133	0.7381	0.5691	55.10	32.00	11.85M
DANet	0.5235	0.3412	0.4971	0.2344	0.2613	0.4219	0.3799	42.86	26.00	6.67M
MedT	0.4786	0.4814	0.4171	0.2374	0.1292	0.6592	0.4005	42.86	16.00	1.41M
ANORAK	0.7463	0.5863	0.6310	0.4966	0.4430	0.7170	0.6034	65.31	60.00	4.10M

Extended Data Fig. 3 | Segmentation performance and intra- and inter-comparisons. a. Segmentations generated by AI at low-power and high-power resolutions, deposited in 10.6084/m9.figshare.24599796. **b.** Comparison of segmentation and prediction performance for ablation experiments. **c.** Comparison of segmentation and prediction performance with other methods.



a

Path 2 \ Path 1	Lep.	Pap.	Aci.	Cri.	Mic.	Sol.
Lep.	48.9	5.9	8.8	0	0	0.4
Pap.	5.3	38.5	11.7	4.5	9.5	3.5
Aci.	29.8	26.7	44.9	6	19	5.1
Cri.	5.3	4.4	21.5	52.2	4.8	17.3
Mic.	9.6	20.7	6.2	4.5	66.7	2
Sol.	1.1	3.7	6.9	32.8	0	71.7

Path 3 \ Path 1	Lep.	Pap.	Aci.	Cri.	Mic.	Sol.
Lep.	64.9	6.7	6.9	1.5	9.5	1.6
Pap.	2.1	53.3	2.6	1.5	4.8	0.4
Aci.	29.8	33.3	85	13.4	47.6	11
Cri.	2.1	0.7	2.6	74.6	0	7.1
Mic.	1.1	3	0.4	1.5	38.1	0.4
Sol.	0	3	2.6	7.5	0	79.5

Path 3 \ Path 2	Lep.	Pap.	Aci.	Cri.	Mic.	Sol.
Lep.	55.7	8.7	11.5	4	10.5	2.2
Pap.	5.1	41.7	9.6	0	21.1	0.4
Aci.	39.2	36.9	75	47.3	47.4	9.2
Cri.	0	1	1	25.3	2.6	15.3
Mic.	0	2.9	0.5	0	14.5	0.4
Sol.	0	8.7	2.4	23.3	3.9	72.5

b

%	Path 1	Path 2	Path 3
Path 1	100	53.49	74.08
Path 2		100	54.20
Path 3			100

	Path 1	Path 2	Path 3
Path 1	1	0.41	0.65
Path 2		1	0.41
Path 3			1

c

%	Path 1	Path 2	Path 3
Path 1	100	71.95	82.01
Path 2		100	73.25
Path 3			100

	Path 1	Path 2	Path 3
Path 1	1	0.40	0.66
Path 2		1	0.41
Path 3			1

e

Path 2 \ Path 1	Grade1	Grade2	Grade3
Grade1	46.8	9.6	1.2
Grade2	27.3	32.6	1.6
Grade3	26	57.8	97.2

Path 3 \ Path 1	Grade1	Grade2	Grade3
Grade1	64.9	5.9	1.2
Grade2	29.9	69.3	7.2
Grade3	5.2	24.8	91.6

Path 3 \ Path 2	Grade1	Grade2	Grade3
Grade1	48.5	13.7	3.5
Grade2	42.6	70.9	20.3
Grade3	8.8	15.4	76.2

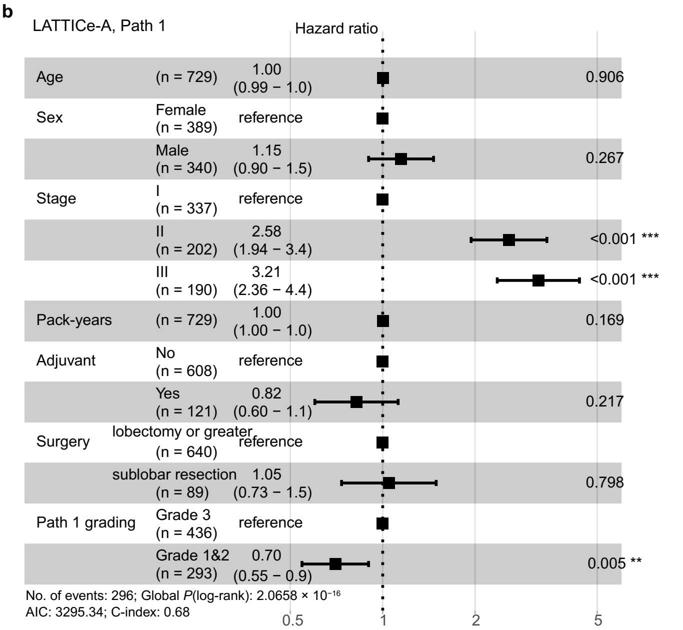
Extended Data Fig. 4 | Inter-pathologists comparison for predominant pattern and IASLC grading in LATTICe-A. a. Interobserver agreement of each pattern. **b, c.** Interobserver agreement of predominant pattern and IASLC grading at tumor level. **d.** Growth pattern intra-tumoral heterogeneity substantially contributed to the discrepancy between pathologists (n = 845 each, P1 < 2.22×10^{-16}, P2 = 4.323×10^{-13}, P3 = 1.589×10^{-15}). P value was calculated using a two-sided Wilcoxon rank-sum test and not adjusted for the multiple comparisons. The median value is indicated by a thick horizontal line; the first and third quartiles are represented by box edges; whiskers indicate 1.5 times interquartile range. **e.** Interobserver agreement of each grade.

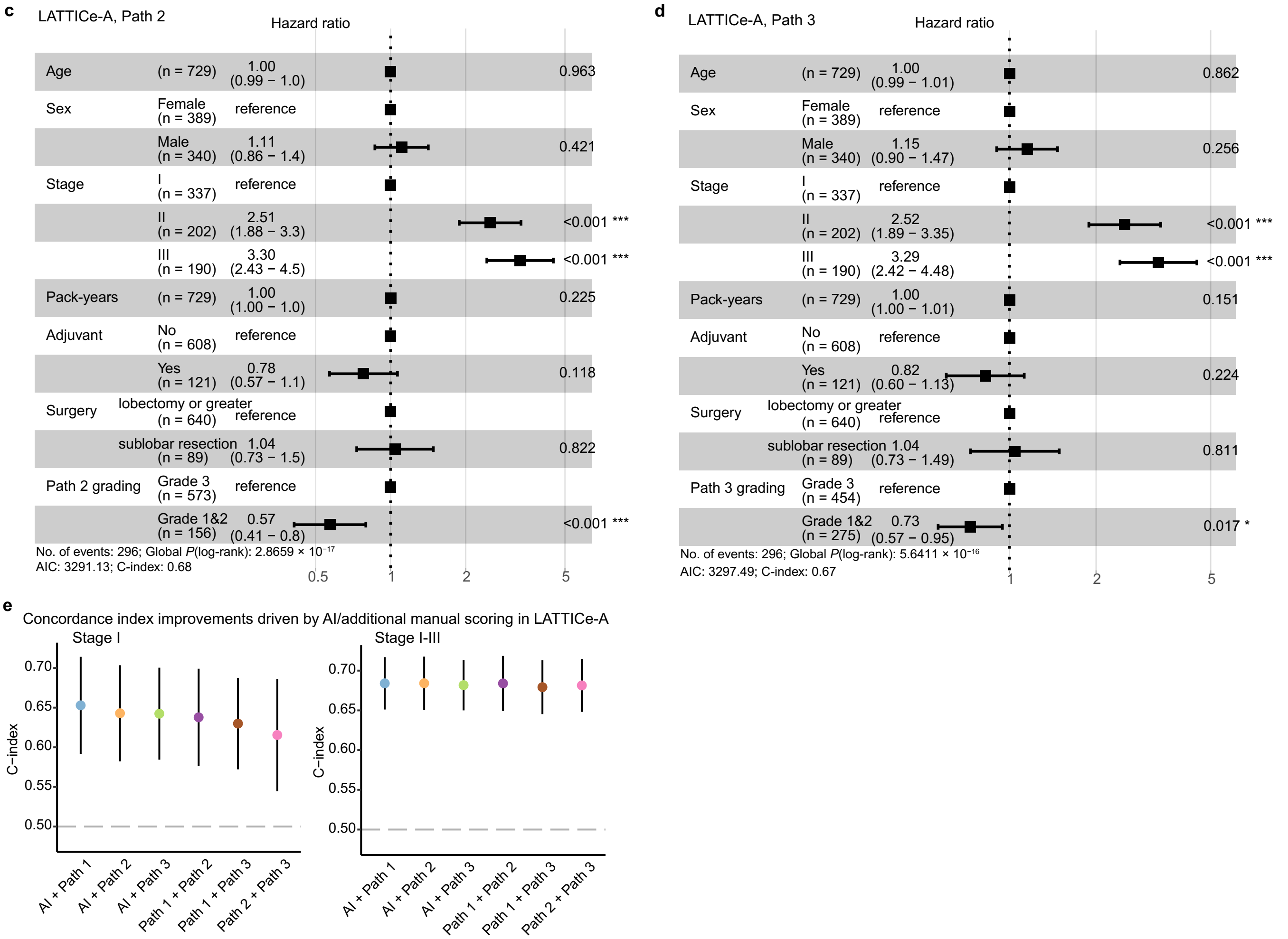
a Pair-wise comparison of AI grades in univariable and multivariable Cox proportional hazards models (DFS)

TRACERx 421	Univariable	Multivariable - Stage[1]	Multivariable - Size[2]
Grade 1&2 vs. Grade 3	HR = 0.48 95% CI = 0.30-0.78	HR = 0.51 95% CI = 0.31-0.85	HR = 0.48 95% CI = 0.29-0.79
Grade 1 vs. Grade 3	HR = 0.31 95% CI = 0.097-1.00	HR = 0.45 95% CI = 0.13-1.51	HR = 0.38 95% CI = 0.12-1.26
Grade 2 vs. Grade 3	HR = 0.53 95% CI = 0.32-0.88	HR = 0.52 95% CI = 0.31-0.89	HR = 0.50 95% CI = 0.29-0.84
Grade 1 vs. Grade 2	HR = 0.59 95% CI = 0.17-2.02	HR = 0.85 95% CI = 0.24-3.02	HR = 0.77 95% CI = 0.22-2.69
Overall *P*-value	*P* = 0.011	*P* = 0.033	*P* = 0.014

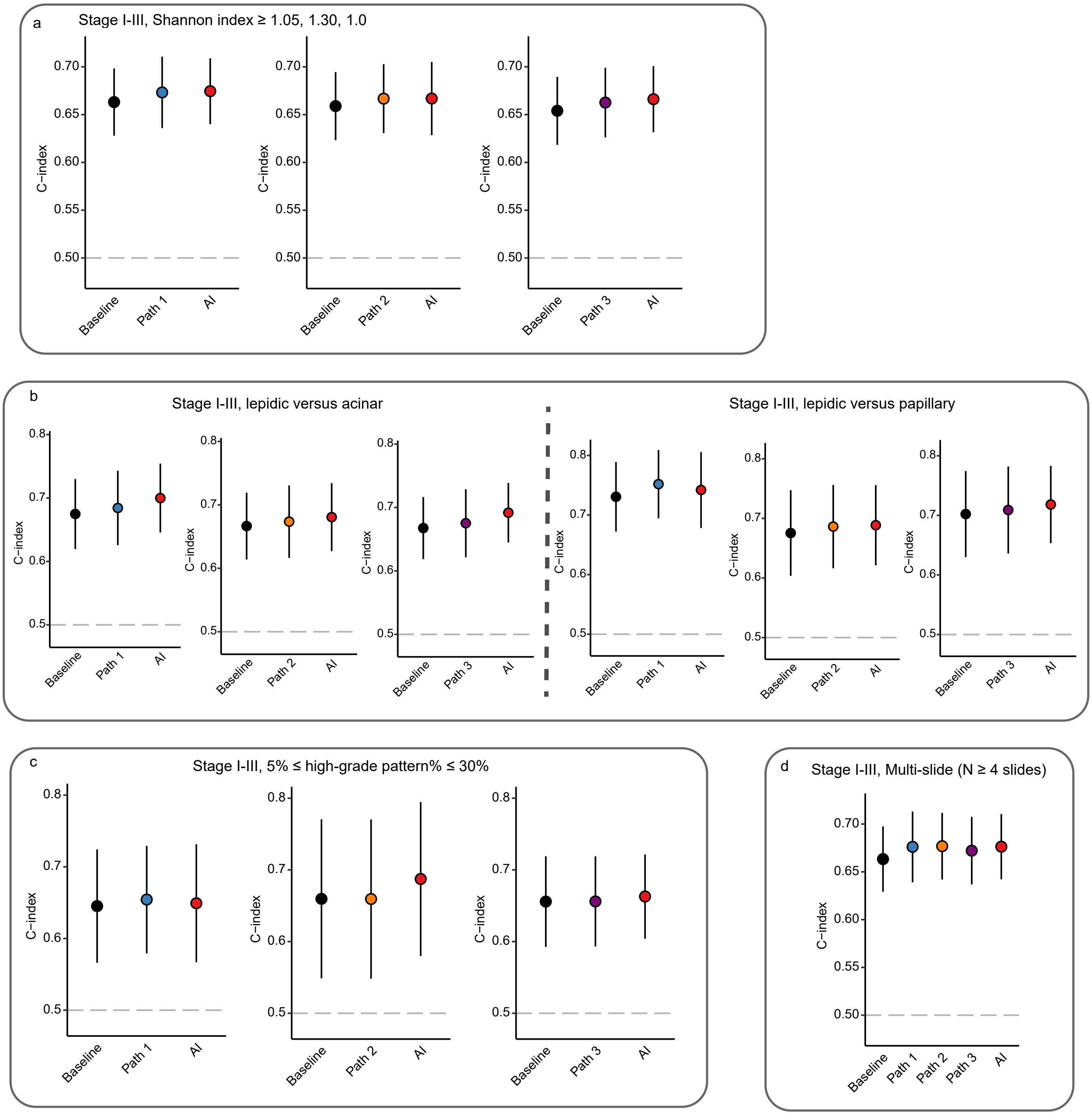
LATTICe-A	Univariable	Multivariable - Stage[1]	Multivariable - Size[2]
Grade 1&2 vs. Grade 3	HR = 0.53 95% CI = 0.42-0.68	HR = 0.64 95% CI = 0.49-0.84	HR = 0.64 95% CI = 0.49-0.84
Grade 1 vs. Grade 3	HR = 0.40 95% CI = 0.24-0.67	HR = 0.52 95% CI = 0.28-0.97	HR = 0.44 95% CI = 0.23-0.83
Grade 2 vs. Grade 3	HR = 0.57 95% CI = 0.45-0.74	HR = 0.66 95% CI = 0.50-0.88	HR = 0.69 95% CI = 0.52-0.91
Grade 1 vs. Grade 2	HR = 0.70 95% CI = 0.40-1.21	HR = 0.79 95% CI = 0.41-1.50	HR = 0.64 95% CI = 0.33-1.24
Overall *P*-value	*P* = 7.81×10^{-7}	*P* = 0.004	*P* = 0.003

[1] Adjusted for age, sex, **stage**, pack-years, surgery type, and adjuvant therapy.
[2] Adjusted for age, sex, **tumour size**, pack-years, surgery type, and adjuvant therapy.



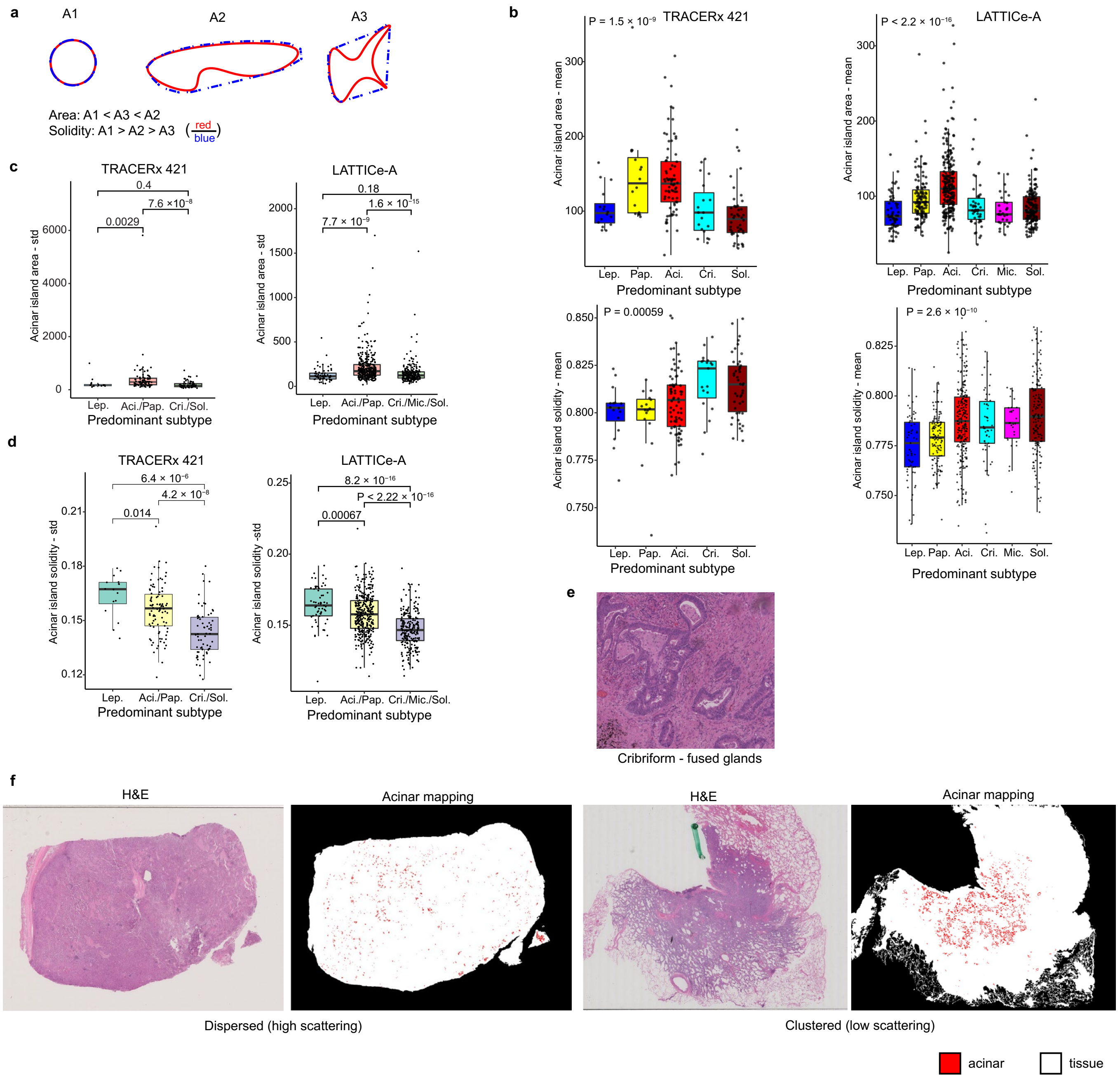


Extended Data Fig. 5 | Survival analyses of AI and pathological gradings. **a**. Pair-wise comparison of AI grades in univariable and multivariable Cox proportional hazards models. **b**–**d**. Multivariable Cox regression analyses showing pathological gradings were independent of age, sex, tumor stage, smoking pack-years, adjuvant therapy, type of surgery in LATTICe-A (P1 = 0.00524, P2 = 0.000913, P3 = 0.0169). HRs of each variable with 95% confidence intervals are shown on the horizontal axis; P value was derived with Wald test. Asterisks indicate: *P < 0.05, **P < 0.01, ***P < 0.001. **e**. Comparison of improvements driven by AI and additional manual scoring for stage I (n = 337) and stage I-III (n = 729) tumors in LATTICe-A, where models included age, sex, tumor stage and gradings from AI or/and pathologists. C-indexes with 95% confidence intervals are shown on the vertical axis.

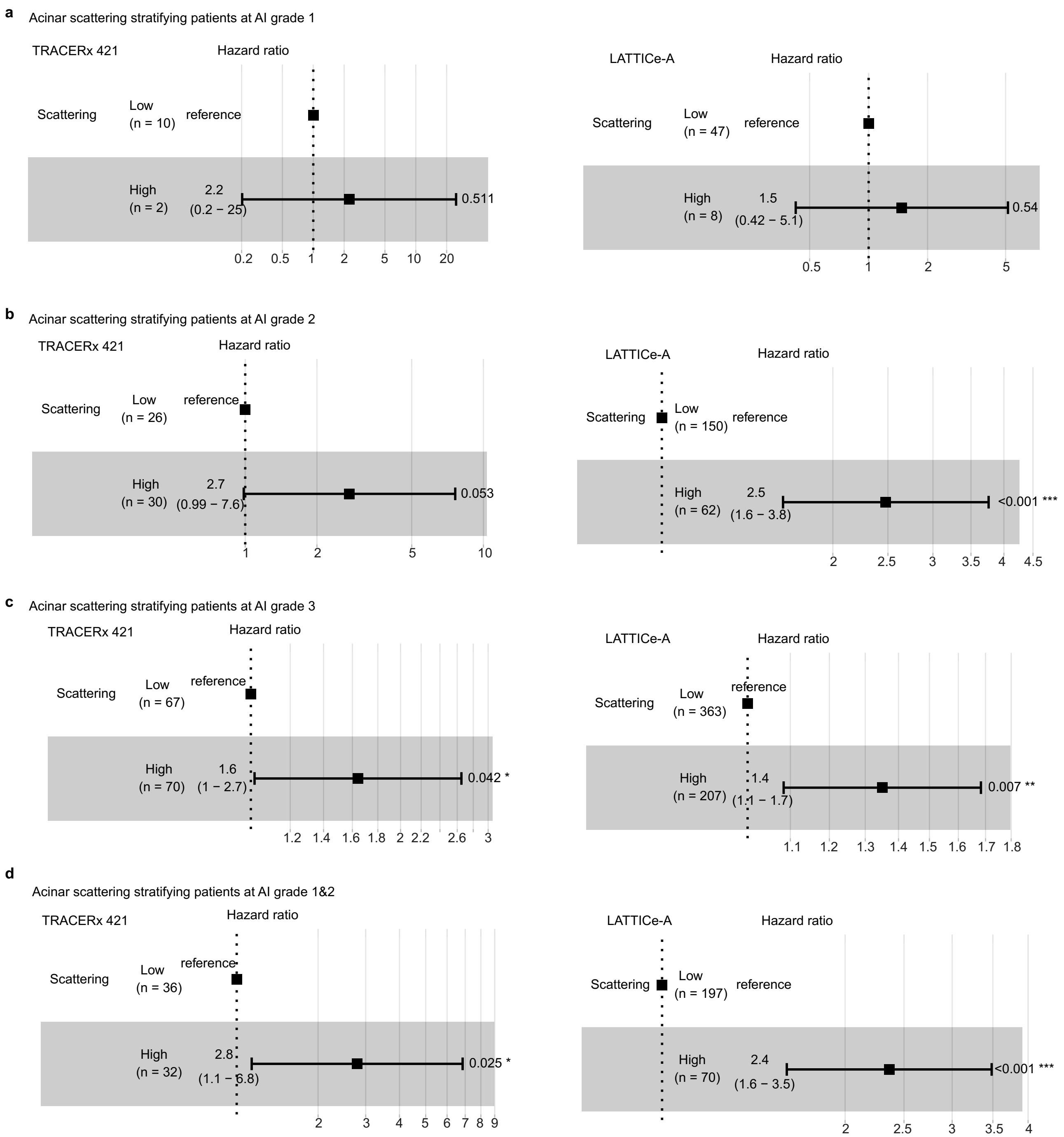


Extended Data Fig. 6 | Assistance of AI in grading challenging scenarios for stage I-III tumors in LATTICe-A. a. Challenging scenario 1, tumors with highly diversified growth patterns indicated by the Shannon diversity index (n1 = 363, n2 = 361, n3 = 390). **b**. Challenging scenario 2, differentiation between lepidic- and acinar-predominant tumors (n1 = 274, n2 = 222, n3 = 340), and between lepidic- and papillary-predominant tumors (n1 = 162, n2 = 134, n3 = 137). **c**. Challenging scenario 3, tumors with high-grade patterns between 5% and 30% (n1 = 162, n2 = 117, n3 = 252). **d**. Challenging scenario 4, tumors with no less than 4 slides (n = 551). C-indexes of each variable with 95% confidence intervals are shown on the vertical axis.

https://doi.org/10.1038/s43018-023-00694-w



Extended Data Fig. 7 | Morphological and spatial analyses of acinar island. **a**. Acinar morphological feature measures, area and solidity index. **b**. Acinar islands are morphologically different among tumors with different predominant patterns (TRACERx 421, P = 1.493 × 10^{-9} and P = 0.0005932, n = 173; LATTICe-A, P < 2.22 × 10^{-16} and P = 2.626 × 10^{-10}, n = 654). P value was calculated using a one-way Kruskal-Wallis rank-sum test and not adjusted for the multiple comparisons. **c**. Acinar island areas were less varied in lepidic-predominant (TRACERx 421, P = 0.002889, n = 108; LATTICe-A, P = 7.743 × 10^{-9}, n = 420) and high-grade-predominant (TRACERx 421, P = 7.617 × 10^{-8}, n = 157; LATTICe-A, P = 1.611 × 10^{-15}, n = 593) tumors than acinar- and papillary-predominant tumors. **d**. Acinar island shapes were less varied in high-grade-predominant tumors than lepidic predominant tumors (TRACERx 421, P = 6.374 × 10^{-6}, n = 81; LATTICe-A, P = 8.184 × 10^{-16}, n = 295). **b**-**d**. Each point is a tumor, y axis is the standard deviation of the area or solidity index for all the individual acinar islands within a tumor. The median value is indicated by a thick horizontal line; the first and third quartiles are represented by box edges; whiskers indicate 1.5 times interquartile range. **c**-**d**. P value was calculated using a two-sided Wilcoxon rank-sum test and not adjusted for the multiple comparisons. **e**. Example illustrating the transition from acinar to cribriform. **f**. Examples of high and low acinar scattering inferred from H&E images with the AI method, deposited in 10.6084/m9.figshare.24599796.



Extended Data Fig. 8 | Acinar scattering stratifying subgroups of AI grading. **a**. Acinar scattering stratifying patients at AI grade 1 (TRACERx 421 P = 0.5112, n = 12; LATTICe-A P = 0.5397, n = 55). **b**. Acinar scattering stratifying patients at AI grade 2 (TRACERx 421 P = 0.0533, n = 56; LATTICe-A P = 1.947×10^{-5}, n = 212). **c**. Acinar scattering stratifying patients at AI grade 3 (TRACERx 421 P = 0.04235, n = 137; LATTICe-A P = 0.007446, n = 570). **d**. Acinar scattering stratifying patients at AI grades 1&2 (TRACERx 421 P = 0.02517, n = 68; LATTICe-A P = 1.387×10^{-5}, n = 267). HRs of each variable with 95% confidence intervals are shown on the horizontal axis. P value was derived with Wald test, and not adjusted for multiple comparisons. Asterisks indicate: *P < 0.05, **P < 0.01, ***P < 0.001.

nature portfolio

Corresponding author(s): Yinyin Yuan, John Le Quesne, David A. Moore

Last updated by author(s): Nov 21, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Main source of data collection is microscopic H&E scanning. We used the NanoZoomer Digital Pathology System version 3.1.7 (Hamamatsu, Japan) for digitalizing H&E slides.
Data analysis	AI pipeline generating growth pattern mask is available at https://github.com/xi11/AIgrading Python v3.8 h5py v2.10.0 keras v2.4.3 numpy v1.20.3 opencv-python v4.5.3.56 pandas v1.3.2 pillow v8.3.1 scipy v1.7.1 tensorflow-gpu v2.2.0 for training or inference tensorflow v2.2.0 can also be used for inference Statistical analysis code is available at https://github.com/xi11/AIgrading R v4.1.2 stats v4.1.2 tidyverse v2.0.0 irr v0.84.1 caret v6.0-93

tidyr v1.3.0
survminer v0.4.9
survival v3.2-13
ggplot2 v3.4.1
ggpubr v0.5.0
survC1 v1.0-3
car v3.0-12
RColorBrewer v1.1-3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The training dataset consisting of annotations on small image tiles are deposited in Zenodo (doi.org/10.5281/zenodo.10016027). Previously published image data that were re-analysed in this study can be requested from https://bmirds.github.io/LungCancer/. The human lung adenocarcinoma diagnostic slide images were derived from the TCGA Research Network: https://portal.gdc.cancer.gov/. Source data for Figures 2, 3, 5b, 5c, 5d, 5e, 5g, 5h, 5i and Extended Data Figures 2, 3a, 5a, 7b, 7c, 7d, 7f, except for clinical data of LATTICe-A cohort, have been provided as Source Data files. Images generated by the AI model in Figure 2a, Extended Data Figures 2, 3a and 7f are deposited in 10.6084/m9.figshare.24599796. For the TRACERx study, all of the scanned diagnostic histological images have a study number label embedded in the file which prevents complete anonymisation. These images cannot therefore be shared, in line with the ethical approval of the study. Request for access to the TRACERx dataset for academic non commercial research purposes can be submitted through the Cancer Research UK and UCL Cancer Trials Centre (ctc.tracerx@ucl.ac.uk), and subject to review of a project proposal that will be evaluated by a TRACERx data access committee, entering into an appropriate data access agreement and any applicable ethical approvals. The timeframe of response to requests is about 6 months. LATTICe-A study data and materials are currently subject to a material and data transfer agreement between University of Leicester, University of Cambridge and NHS Greater Glasgow and Clyde which includes a restricted access period of 5 years precluding any access by other third parties during this time. After the 5 year restricted access data can be accessed by application to NHS Greater Glasgow and Clyde Biorepository (clare.orange@ggc.scot.nhs.uk; john.lequesne@glasgow.ac.uk) as custodians and the data access request will be reviewed and released under their REC approved tissue bank protocols. Requests are reviewed and approved within 6-8 weeks and will be accompanied by a data sharing agreement detailing conditions and restrictions of use and publication.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender

Sex and gender were not considered in the study design. For TRACERx 421, self-reported sex was collected and used in the analyses, denoted as 'Sex'. For LATTICe-A, sex assigned at birth was collected and used in the analyses, denoted as 'Sex'.

Population characteristics

Please note that the study started recruiting patients in 2016, when TNM version 7 was standard of care. The up-to-date inclusion/exclusion criteria now utilizes TNM version 8.

TRACERx inclusion and exclusion criteria

Inclusion Criteria:
_Written Informed consent
_Patients ≥18 years of age, with early stage I-IIIB disease (according to TNM 8th edition) who are eligible for primary surgery.
_Histopathologically confirmed NSCLC, or a strong suspicion of cancer on lung imaging necessitating surgery (e.g. diagnosis determined from frozen section in theatre)
_Primary surgery in keeping with NICE guidelines planned
_Agreement to be followed up at a TRACERx site
_Performance status 0 or 1
_Minimum tumor diameter at least 15mm to allow for sampling of at least two tumour regions (if 15mm, a high likelihood of nodal involvement on pre-operative imaging required to meet eligibility according to stage, i.e. T1N1-3)

Exclusion Criteria:
_Any other* malignancy diagnosed or relapsed at any time, which is currently being treated (including by hormonal therapy).
_Any other* current malignancy or malignancy diagnosed or relapsed within the past 3 years**.
*Exceptions are: non-melanomatous skin cancer, stage 0 melanoma in situ, and in situ cervical cancer
**An exception will be made for malignancies diagnosed or relapsed more than 2, but less than 3, years ago only if a pre-operative biopsy of the lung lesion has confirmed a diagnosis of NSCLC.
_Psychological condition that would preclude informed consent
_Treatment with neo-adjuvant therapy for current lung malignancy deemed necessary
_Post-surgery stage IV
_Known Human Immunodeficiency Virus (HIV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV) or syphilis infection.
_Sufficient tissue, i.e. a minimum of two tumor regions, is unlikely to be obtained for the study based on pre-operative imaging

Patient ineligibility following registration
_There is insufficient tissue
_The patient is unable to comply with protocol requirements
_There is a change in histology from NSCLC following surgery, or NSCLC is not confirmed during or after surgery.
_Change in staging to IIIC or IV following surgery
_The operative criteria are not met (e.g. incomplete resection with macroscopic residual tumors (R2)). Patients with microscopic residual tumors (R1) are eligible and should remain in the study
_Adjuvant therapy other than platinum-based chemotherapy and/or radiotherapy is administered.

The external validation cohort was obtained from the Leicester Archival Thoracic Tumor Investigatory Cohort-Adenocarcinoma (LATTICe-A) study which consists of 845 University Hospitals of Leicester (UHL) Trust patients who underwent surgical treatment with curative intent for primary invasive non-mucinous lung adenocarcinoma. 401 were men and 444 were women with a mean age of 67.66. Most clinical data (age, sex, adjuvant therapy status and time to recurrence or death) were available for all patients, with complete pathological stage for 729 and smoking history for 742. LATTICe-A patients were initially identified using diagnostic histopathology reports from surgical specimens within UHL's local histopathology database.
Inclusion criteria:
• Surgical resection procedures included are wedge resection; lobectomy; bilobectomy; segmentectomy; and pneumonectomy.
• Patients diagnosed with synchronous primary lung tumours, including non-adenocarcinoma tumours were considered eligible for inclusion
• Patients >= 18 years of age.
Exclusion criteria:
• Patients must not have been diagnosed with any other lung tumour in the previous five years before surgical treatment for the primary lung adenocarcinoma.
• Patients who have been diagnosed with metastatic lung adenocarcinoma.
• Lung adenocarcinoma tumours diagnosed as a recurrence of a previous primary lung adenocarcinoma.
• Patients who have only had biopsy surgery performed i.e. core biopsies/EBUS.
• Patients who have been diagnosed with combined tumour types e.g. adenosquamous carcinoma or combined adenocarcinoma/high grade neuroendocrine carcinoma.
• Patients who have been diagnosed with undifferentiated non-small cell lung tumours.

Recruitment

Patients seen with a new diagnosis of lung cancer in lung cancer units across the United Kingdom, according to the eligibility criteria above, were recruited. No selection bias has been identified to date.
For TRACERx, clinical and pathological data is collected from patients during study follow up - this period is a minimum of five years. Data collection is overseen by the sponsor of the study (Cancer Research UK & UCL Cancer Trials Centre) and takes place in 19 hospitals across the United Kingdom. A centralised database called MACRO is used for this purpose. Recruitment started in April 2014 and is still ongoing in London and Manchester. Survival data last updated in 15 June 2021. LATTICe-A, the Leicester Archival Thoracic Tumor Investigatory Cohort – Adenocarcinoma, is a continuous retrospective series of resected primary LUAD tumors from a single surgical center, University Hospitals of Leicester (UHL) Trust, between 1998 and 2014.
In TRACERx, disease-free survival (DFS) was measured from the time of study registration to date of first lung recurrence or death from any cause. Patients who didn't have these events were censored at the date last known to be alive (including patients who developed a new primary tumor that has been shown biologically not to be linked to the initial primary lung tumor). Overall survival was measured from the time of study registration to date of death from any cause.
In the LATTICE-A cohort, recurrence data were obtained from examination of patient records, notably paper notes and radiological databases, to identify the date of radiologically or biopsy-confirmed recurrence. Cancer-specific death is determined by the presence of lung cancer in the cause of death in the death certificate. Overall survival refers to the date of death.

Ethics oversight

The study was approved by the NRES Committee London with the following details:
Study title: TRAcking non small cell lung Cancer Evolution through therapy (Rx)
REC reference: 13/LO/1546
Protocol number: UCL/12/0279
IRAS project ID: 138871
Study protocol: https://clinicaltrials.gov/ct2/show/NCT01888601

The LATTICe-A cohort was ethically approved by an NHS research ethics committee (ref. 14/EM/1159), more information can be found at: https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/researchsummaries/characterisation-of-thoracic-malignancies-using-archival-human-tissue/

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Primary tumours diagnosed as invasive non-mucinous lung adenocarcinoma from 206 patients were analysed for TRACERx 421. The sample size is similar to the one reported in a previous publication (PMID: 37045996, 37046096) and subject to available diagnostic slides. Primary tumours diagnosed as invasive non-mucinous lung adenocarcinoma from 845 patients were analysed for LATTICe-A cohort. The sample size is similar to the one reported in a previous publication (PMID: 32461698). Primary tumours diagnosed as invasive non-mucinous lung adenocarcinoma from 178 patients were analysed for TCGA, with a similar samples size reported in a previous publication (PMID: 25079552) and subject to available diagnostic slides. The sample size, 143, in DHMC is as same as the one reported in a previous publication (PMID: 30833650). The objective of this study is to validate the performance of AI model against manual scorings, compared with similar studies (PMID: 30833650, 30728398), 1372 cases from four independent cohorts are sufficient.
Data exclusions	Please see study inclusion/exclusion criteria above. Additionally, tumours diagnosed as mucinous lung adenocarcinoma, adenocarcinoma in situ, minimally invasive adenocarcinomas and other variants were excluded.
Replication	The AI model developed for TRACERx 100 was applied directly to TRACERx 421, LATTICe-A, TCGA and DHMC cohorts, achieving a degree of agreement with human pathologists equivalent to inter-pathologist agreement and allowing the replication of the prognostic value of AI grading in TRACERx 421 and LATTICe-A cohorts.
Randomization	Randomization is not relevant as this is an observational study.
Blinding	Blinding is not relevant as this is an observational study. Patients were not allocated to any interventions and they were followed up and assessed as per routine practice. No results from this study are reported back to patients, so there is no likelihood of people changing their behaviors based on these findings. The deep learning model was training without knowing the outcome of patients, which represents a form of blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a | Involved in the study

- ☒ □ Antibodies
- ☒ □ Eukaryotic cell lines
- ☒ □ Palaeontology and archaeology
- ☒ □ Animals and other organisms
- ☒ □ Clinical data
- ☒ □ Dual use research of concern

Methods

n/a | Involved in the study

- ☒ □ ChIP-seq
- ☒ □ Flow cytometry
- ☒ □ MRI-based neuroimaging