



Quantifying Early-Stage Lung Adenocarcinoma Progression with a Radiomic Trajectory



Zhen-Bin Qiu^{1,2,3,11}, Jiaqi Li^{4,5,11}, Shihua Dou^{6,7,11}, Qiuchen Meng⁴, Meng-Min Wang^{1,3}, Hong-Ji Li^{1,2,3}, Chao Zhang^{1,3}, Hongsheng Xie⁶, Ben-Yuan Jiang^{1,3}, Jun-Tao Lin^{1,3}, Jia-Tao Zhang^{1,3}, Fang-Ping Xu⁸, Jin-Hai Yan⁸, Lei Wei⁴, Yi-Long Wu^{1,3}, Haibo Wang⁹, Lin Yang⁶✉, Xuegong Zhang^{4,10}✉ & Wen-Zhao Zhong^{1,2,3}✉

Determining tumor progression status is critical for early-stage lung adenocarcinoma (esLUAD) diagnosis and treatment, yet histopathology-based grading often overlooks heterogeneity within grades. We propose RadioTrace, a deep contrastive learning framework integrating radiomic and pathological information to learn a radiomic trajectory for quantifying esLUAD progression. Across four multi-institutional cohorts, RadioTrace well predicted tumor phenotypes including spread through air spaces (STAS) and lymph node metastasis (LNM). Survival analyses demonstrated it as an independent prognostic factor (log-rank test $p < 0.004$ across all cohorts). Within the same pathological grade, it revealed significant survival heterogeneity ($p < 0.02$ across all cohorts), underscoring the limitations of current grading criteria. Genomic and transcriptomic analyses confirmed associations with progression-related molecular features. Longitudinal analysis of patients with multiple CT follow-ups further showed consistency with continuous progression. These findings demonstrate that RadioTrace enables quantitative, interpretable assessment of esLUAD progression, providing insights beyond histopathology and assisting clinical decision-making.

Lung cancer is the leading cause of cancer-related deaths in the world, and lung adenocarcinoma (LUAD) is the most common subtype¹. LUAD progresses from precursor glandular lesion to malignant disease, during which it gradually gains the abilities of invasiveness and metastasis². Consequently, early detection of LUAD is crucial for effective treatment³. Early-stage non-mucinous LUAD (esLUAD) is generally acknowledged to progress undergoing four pathological stages: atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC)⁴. Recently,

pathologists further subdivided IAC into grades I-III as per the proportion of different histological subtypes⁵. These grading criteria are routinely used in clinical diagnosis and facilitate precise treatment selection^{6,7}.

However, the reliance on histopathological patterns for tumor classification presents challenges. This method is an invasive approach and cannot provide a comprehensive assessment of tumors due to sampling bias⁸. Moreover, the classification of histological subtypes is inherently subjective, leading to significant interobserver variability among pathologists. Studies have shown relatively low κ -values for overall agreement,

¹Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China. ²School of Medicine, South China University of Technology, Guangzhou, China. ³Guangdong Provincial Key Laboratory of Translational Medicine in Lung Cancer, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China. ⁴Bioinformatics Division, BNRIST and MOE Key Lab of Bioinformatics, Department of Automation, Tsinghua University, Beijing, China. ⁵China Mobile Research Institute, Beijing, China. ⁶Department of Thoracic Surgery, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University, the First Affiliated Hospital of South University of Science and Technology of China, Shenzhen Institute of Respiratory Diseases), Shenzhen, China. ⁷Department of Thoracic Surgery, the First Affiliated Hospital of Hainan Medical University, Hainan Province Clinical Medical Center of Respiratory Disease, Haikou, China. ⁸Department of Pathology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China. ⁹Clinical Trial Unit, Precision Medicine Institute, First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China. ¹⁰School of Medicine, Tsinghua University, Beijing, China. ¹¹These authors contributed equally: Zhen-Bin Qiu, Jiaqi Li, Shihua Dou. ✉e-mail: 13798314779@163.com; zhangxg@tsinghua.edu.cn; syzhongwenzhao@scut.edu.cn

ranging from 0.237 to 0.646, which may compromise the accuracy of patient prognosis assessments^{9–12}. Additionally, tumor progression is a continuous process, but the conventional stepped pathological grading system may overlook differences within the same grade and result in varying responses to the same treatment strategy based on pathological diagnosis. For instance, Park BJ et al. suggested to subdivide grade II-IAC into II-a and II-b based on the proportion of lepidic component¹³. Variations such as the morphology and spatial distribution of acinar islands within tumors, which are not accounted for in current LUAD grading systems, were also found to significantly affect patient prognosis¹⁴. Besides, the boundary differences between grades may be exaggerated in scenarios. Our previous study indicated that different pathological grades with similar radiological features may have comparable recurrence risks¹⁵. These observations underscore the urgent need for a continuous measurement of tumor progression to enhance the precision of diagnosis and treatment of esLUAD¹⁶.

Computed tomography (CT) is a non-invasive technology commonly used in the detection and diagnosis of lung cancer¹⁶. CT images provide a comprehensive view of tumor tissue by transforming the composition and spatial distribution of cells into various gray levels of pixels¹⁷. With the development of radiomics, scientists have proposed many computational methods, including deep learning-based methods, to discriminate the pathological grades of esLUAD^{18–20}. These studies demonstrated the feasibility of characterizing esLUAD progression with CT images, but they didn't attempt to obtain additional information of tumor progression beyond pathological grades. Such approaches may limit the model to extracting information that is not related to pathology but is closely related to tumor progression, and cannot provide continuous quantification of the progression. Quantifying biological processes with a continuous trajectory have been widely implemented in the field of single-cell transcriptomic studies²¹, and similar idea has emerges in histopathology studies²², but to the best of our knowledge, there have been no existing approaches that explicitly build a continuous “radiomic trajectory” to decipher tumor progression with pathological and/or radiomic information.

In this study, we proposed a new strategy to quantitatively measure the progression of esLUAD. We integrated the pathological and radiomic information of tumors using a deep contrastive learning framework, and developed a radiomic esLUAD progression trajectory named RadioTrace. We validated the effectiveness and consistency of RadioTrace on multiple patient cohorts from different institutes. We found that the pseudo-progression score (PPS) derived from RadioTrace is able to predict the appearance of the spread through air spaces (STAS), the lymph node metastasis (LNM), and patient's survival conditions in LUAD. Even within the grade II IAC, tumors with different PPSs were significantly different in phenotypes and prognoses. We also performed radiogenomic analysis and found that RadioTrace is associated with the tumor-progression-related gene alterations and differentially expressed genes. In addition, RadioTrace was shown to be highly consistent with longitudinal esLUAD progression records of the same patient. The results showed that combining multi-scale tumor characteristics is able to quantify the progression procedure of esLUAD and uncover more detailed differences than the pathological grades. The developed RadioTrace can assist the identification of esLUAD progression status in the clinical diagnosis. We developed the corresponding training and inference methods of RadioTrace as a Python package freely available for academic use at <https://pypi.org/project/RadioTrace>.

Results

Patient cohorts

This study is a multi-center study which includes four in-hospital cohorts of patients, three from Guangdong Provincial People's Hospital (the GDPH1, GDPH2, and GDPH3 cohort, *Methods*), and one from Shenzhen People's Hospital (the SZPH cohort, *Methods*), whom were treated with surgery only. As Fig. 1a shown, 1843 tumors of 1827 patients were finally enrolled in this study. The GDPH1 cohort consists 641 tumors of 625 patients from 2015–2020, which was used for model training and radiomic trajectory construction. The GDPH2 cohort with 158 tumors of 158 patients from

2012–2014 was used for internal validation. The GDPH3 cohort with 195 tumors of 195 patients was used for multiomics validation. In GDPH 3 cohort, all 195 patients with bulk RNA-seq data of tumors as well as 87 patients of them with panel DNA-NGS data of tumors were used for genetic and transcriptomic association analysis. The SZPH cohort consisting 849 tumors of 849 patients was used for external validation. Most cohorts include patients across five pathological grades ranging from AIS to IAC-III, except for the GDPH2 cohort, which includes only grades IAC-I to IAC-III, and the GDPH3 cohort, which does not include patients with IAC-I. Most patients were diagnosed with esLUAD (in stage T1N0 or stage T2aN0) except that a small group of patients in the SZPH cohort were in stage T1N1 or T1N2. The demographic and clinical information of patients is shown in Table 1. Additionally, we also collected data from Guangdong Provincial People's Hospital as a case study for two patients who were not included in any of the cohorts described above, and each patient had longitudinal CT images, including an initial and 4 follow-up CT scans.

Constructing a radiomic trajectory of esLUAD progression with deep contrastive learning

We proposed a deep contrastive learning framework, based on the assumption that tumors at similar stages of evolution share similar imaging features, to infer the trajectory of tumor evolution using large-scale tumor data. Pathological grading system involves two dimensions: categorical classification and sequential progression. Only the categorical relationship—whether two samples belong to the same pathological category—is used as weak supervision. This strategy helps blur rigid grade boundaries and fully leverages the feature extraction capabilities of a 3D deep convolutional neural network (CNN) to explore both inter-tumor similarities and intra-grade variations. As shown in Fig. 1b, c, the basic idea is to train a deep learning model to obtain an embedding for the CT image of each tumor, where embeddings of tumors with the same pathological grade are encouraged to be close, while embeddings of tumors with different pathological grades are encouraged to be far apart. (*Methods*). Through this process, the CNN model learns to capture imaging features indicative of tumor evolution while incorporating pathological context. We used the GDPH1 cohort as the training data. Details about the model is provided in *Methods*.

After training, we performed principal component analysis (PCA) on the embeddings and visualized the distribution of the samples using the first two principal components (PCs). We then employed the slingshot method²³ to infer a trajectory in the PC space (Fig. 1c) using the AIS cluster as a starting point and took it as a reference of esLUAD progression from AIS to IAC-III (Fig. 2a, *Methods*). We named the trajectory as RadioTrace, as it was built according to the radiomic information of tumors. We defined a pseudo-progression score (PPS) to quantify the positions of tumors on the trajectory (*Methods*), and the PPSs increased gradually with tumor progression, consistent with their pathological grades (Fig. 2b). We applied the algorithm of RadioTrace on GDPH2, GDPH3 and SZPH cohorts for internal or external validation. We found that their radiomic trajectories and PPSs were also well consistent with pathological grades (Fig. 2c–h). It is worth noting that some tumors in the SZPH cohort were in the T1N1 or T1N2 stage, which are more advanced progression statuses than those in the training cohort. Though the trained model was still able to project them to the tail of the RadioTrace (Fig. S1).

We compared RadioTrace with trajectories constructed by pre-defined radiomic features which are widely used in radiomic studies²⁴. We calculated the radiomic features for each tumor in the GDPH1 cohort, including the first-order statistics, shape features, and texture features (*Methods*). We performed dimension reduction on these features by PCA. As shown in Fig. S2, there is no clear direction of tumors from AIS to IAC-III, and no significant difference was shown among tumors with adjacent pathological grades. The result demonstrated the superiority of the deep contrastive learning framework in constructing trajectories reflecting esLUAD progression.

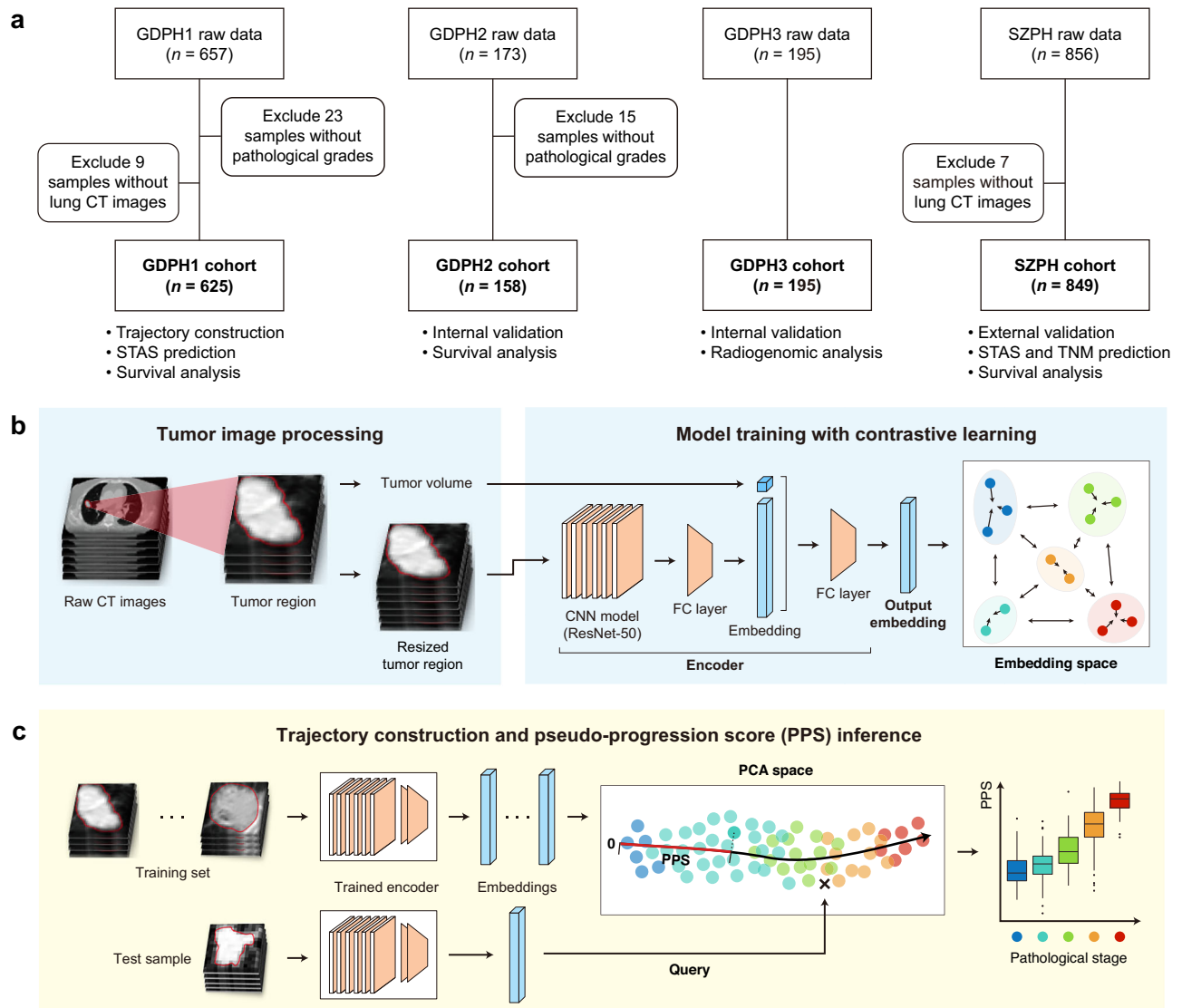


Fig. 1 | Study design and method diagram. **a** Illustration of patient cohorts; **b** workflow of model training with contrastive learning; **c** workflow of trajectory construction and pseudo-progression score (PPS) inference.

RadioTrace associates with phenotypes and prognoses of esLUAD patients

We investigated whether RadioTrace is associated with some key phenotypes of esLUAD patients, including the spread through air spaces (STAS) and the lymph-node metastasis (LNM), both of which can reflect tumor progression and are regarded as important risk factors of patient prognosis²⁵. For STAS, we collected 325 and 521 patients with pathologically validated STAS records in GDPH1 and SZPH cohorts, respectively. We found significant differences between the PPSs of the STAS-present and STAS-absent patients (Wilcoxon rank-sum test, GDPH1 cohort: $p < 0.001$, SZPH cohort: $p < 0.001$; Fig. 3a), while the differences are not significant between patients with STAS in the two cohorts ($p = 0.101$; Fig. 3a). For LNM, we found that PPSs of the LNM-present group were significantly higher than those of the LNM-absent group ($p < 0.001$; Fig. 3b).

We conducted survival analysis on all patients who received lobectomy which is a standard treatment strategy for early-stage lung cancer. We collected the disease-free survival (DFS) time of 260 patients in the GDPH1 cohort, 139 patients in the GDPH2 cohort and 700 patients in the SZPH cohort. We used the x-tile software²⁶ to find the optimal PPS threshold (2.704) for differentiating DFS time in the GDPH1 cohort, and found that patients with lower PPSs showed longer DFS time (log-rank test, $p < 0.001$, HR = 9.699, 95% CI: 2.314–40.65, Fig. 3c). We validated the optimal

threshold on the GDPH2 and SZPH cohort and obtained similar results (GDPH2 cohort: $p = 0.003$, HR = 3.834, 95% CI: 1.456–10.09, Fig. 3d; SZPH cohort: $p < 0.001$, HR = 19.64, 95% CI: 7.083–54.43, Fig. 3e). Then, we conducted univariate and multi-variate analyses using PPSs and some key clinical factors of patients (age, gender, smoking status, EGFR mutation, and tumor diameter). We combined the GDPH1 and GDPH2 cohorts as both the cohorts observed few numbers of DFS events. We found that the PPS is the most significant predictor for patient’s prognosis in both GDPH1/2 and SZPH cohorts (Tables 2–3). All these results showed that RadioTrace can well predict phenotypes and prognoses of esLUAD patients.

RadioTrace reveals heterogeneities within pathological grades

Notably, the range of PPSs within the same pathological grade still varies, suggesting that the phenotypes can be different of tumors within the same grade. Thus, we performed the abovementioned analyses for tumors within each grade, respectively. We found that, in the IAC-II group, PPSs showed significant difference between the STAS-present and STAS-absent patients (Wilcoxon rank-sum test, GDPH1 cohort: $p < 0.001$, SZPH cohort: $p = 0.009$; Fig. 3f), and there was no significant difference between patients with STAS in the two cohorts ($p = 0.292$). The difference of PPSs between the LNM-present and LNM-absent group was also significant ($p < 0.001$, Fig. 3g).

Table 1 | Clinical information of patient cohorts used in this study

Characteristics	GDPH1 (N = 625)	GDPH2 (N = 158)	GDPH3 (N = 195)	SZPH (N = 849)
Source	GDPH	GDPH	GDPH	SZPH
Age (mean ± SD)	57.8 11.8	60.5 11.4	57.7 13.0	52.9 13.0
Gender				
Male	227	84	68	322
Female	398	74	125	527
Smoking				
Yes	98	35	33	103
No	525	123	162	746
T stage				
Tis	85	0	1	113
T1a	137	14	50	379
T1b	215	92	82	273
T1c	121	51	57	65
T2a	47	1	5	19
N stage				
N0	625	158	195	792
N1	0	0	0	20
N2	0	0	0	37
M stage				
M0	625	158	195	849
Pathological grade				
AIS	89	0	1	116
MIA	101	0	43	296
IAC-I	78	32	0	44
IAC-II	329	99	137	343
IAC-III	28	27	14	50
EGFR mutation				
Mutated	373	58	174	72
Wildtype	234	67	21	777
Survival time				
Median DFS (days)	949	2189.5	-	549
Number of events observed (DFS)	9	21	-	27

We also observed significant difference of DFS time within the IAC-II grade in all the three cohorts with the same optimal threshold mentioned above (log-rank test; GDPH1 cohort: $p = 0.002$, HR = 9.405, 95% CI: 1.704–51.93, Fig. 3h; GDPH2 cohort: $p = 0.015$, HR = 3.592, 95% CI: 1.199–10.76, Fig. 3i; SZPH cohort: $p = 0.002$, HR = 6.346, 95% CI: 1.804–22.32, Fig. 3j). The PPS remains to be a significant predictor for DFS (Tables 4–5). The results suggested significant variations exist in tumors at the IAC-II grade, and it is possible to further subtype the IAC-II grade by RadioTrace for better diagnosis and stratification.

We performed similar analyses on the other pathological grades, but we did not find significant difference both for tumor phenotypes and prognoses (Supplementary Notes).

RadioTrace is associated with genetic alterations and progression-related gene expressions

We investigated the relationship between RadioTrace and the genetic and transcriptomic profiles of tumors. We first collected panel Next-Generation Sequencing (NGS) data of tumors from 87 patients in the GDPH3 cohort

(Methods). We divided these samples into a high-PPS group and a low-PPS group by the median PPS. As shown in Fig. S3, the top three genes in terms of mutation frequency are *EGFR*, *TP53*, and *ERBB2*, with frequencies of 68%, 20%, and 16%, respectively. We analyzed the relationship between genes with mutation frequency above 5% and PPS, and found that the PPS of *TP53*, *LRP1B*, and *SMAD4* gene mutants was significantly higher than that of the wild-type, (Wilcoxon rank-sum test, *TP53*: $p < 0.0001$; *LRP1B*: $p < 0.01$; *SMAD4*: $p < 0.05$; Fig. 4a) while others are comparable. It has been reported that the alternations of *TP53*, *LRP1B* and *SMAD4* were associated with the acquisition for invasiveness and the progression of lung cancer^{27–29}.

We also collected bulk RNA-seq data of tumors from 195 patients in the GDPH3 cohort. Similarly, we divided these samples into a high-PPS group and a low-PPS group by the median PPS. We first conducted differential expression (DE) analysis between the two groups. We performed gene set enrichment analysis (GSEA, Fig. 4b) and found that PPSs were significantly positively correlated with many Hallmark gene sets such as G2M Checkpoint (NES 2.38, $p < 0.001$), E2F Target (NES 2.21, $p = 0.001$), mTORC1 Signaling (NES 2.03, $p = 0.014$), Epithelial Mesenchymal Transition (EMT, NES 1.87, $p = 0.021$), and Glycolysis (NES 2.42, $p < 0.001$). All these gene sets are proved to be manifestations of increased malignancy of tumors^{30,31}.

To further uncover the biological correlates of PPS, we built an XGBoost model using the Hallmark gene set enrichment scores as predictors and applied SHAP analysis to evaluate their contributions (Methods, Fig. S4). The top pathways identified were closely linked to tumor progression and largely overlapped with GSEA results, such as Glycolysis, E2F Target, G2M Checkpoint, and Epithelial Mesenchymal Transition. Notably, Glycolysis showed the strongest positive association with PPS, consistent with enhanced glycolytic metabolism in aggressive tumors.

Gene Ontology (GO) enrichment analyses also showed that these two groups are significantly different in terms of cell proliferation and invasiveness (Methods, Tables S1 and S2). In addition, we analyzed the added value of RadioTrace in terms of current pathological grades (Supplementary Notes). The progression-related EMT pathway was uniquely identified by RadioTrace-based DE genes compared with pathology-based DE genes (Figs. 4b, S5). These results showed that tumors on the two ends of RadioTrace are divergent in terms of gene expression. Also, radiomic approaches provide more tumor information as opposed to the commonly-used pathological grades.

We then examined the correlation of the PPS with gene expression levels (Methods). We found that some significantly correlated genes are marker genes for pathological subtypes in esLUAD, such as the positively correlated genes *SPP1*, *MDK*, and *COL1A1*, and the negatively correlated genes *SFTPC* and *AGER*³² (Fig. 4c). We further grouped genes with similar expression patterns into clusters (i.e., gene modules) using the weighted graph co-expression network analysis (WGCNA) package³³, and explored their correlations with PPSs (Methods). The significantly and highly correlated gene modules were regarded as the transcriptomic signatures of RadioTrace (Fig. 4d). We leveraged four public datasets^{34–37} to study the relationship between these transcriptomic signatures and survival conditions of patients (Methods). In each dataset, we divided patients into two groups by the median expression of each gene module and evaluated the significance of patient's prognosis between the two groups. As shown in Fig. 4e, we observed that the gene module “MEpurple” is able to distinguish the overall survival (OS) (log-rank test, Tang et al., $p < 0.001$; Okayama et al., $p < 0.001$; Rousseaux et al., $p < 0.001$, Shedden et al., $p = 0.001$) and DFS (Okayama et al., $p < 0.001$; Rousseaux et al., $p < 0.001$; Shedden et al., $p < 0.001$; no DFS records for the Tang et al. dataset), while the other gene modules were not. We performed KEGG pathway enrichment of genes in the gene module “MEpurple” and found that these genes exhibited predominant enrichment in the Cell Cycle pathway, alongside several other pathways associated with cellular proliferation (Table S3).

We also evaluated the genetic and transcriptomic differences related to RadioTrace within the IAC-II grade. We selected 134 patients with grade IAC-II from the GDPH3 cohort, and divided them into two groups

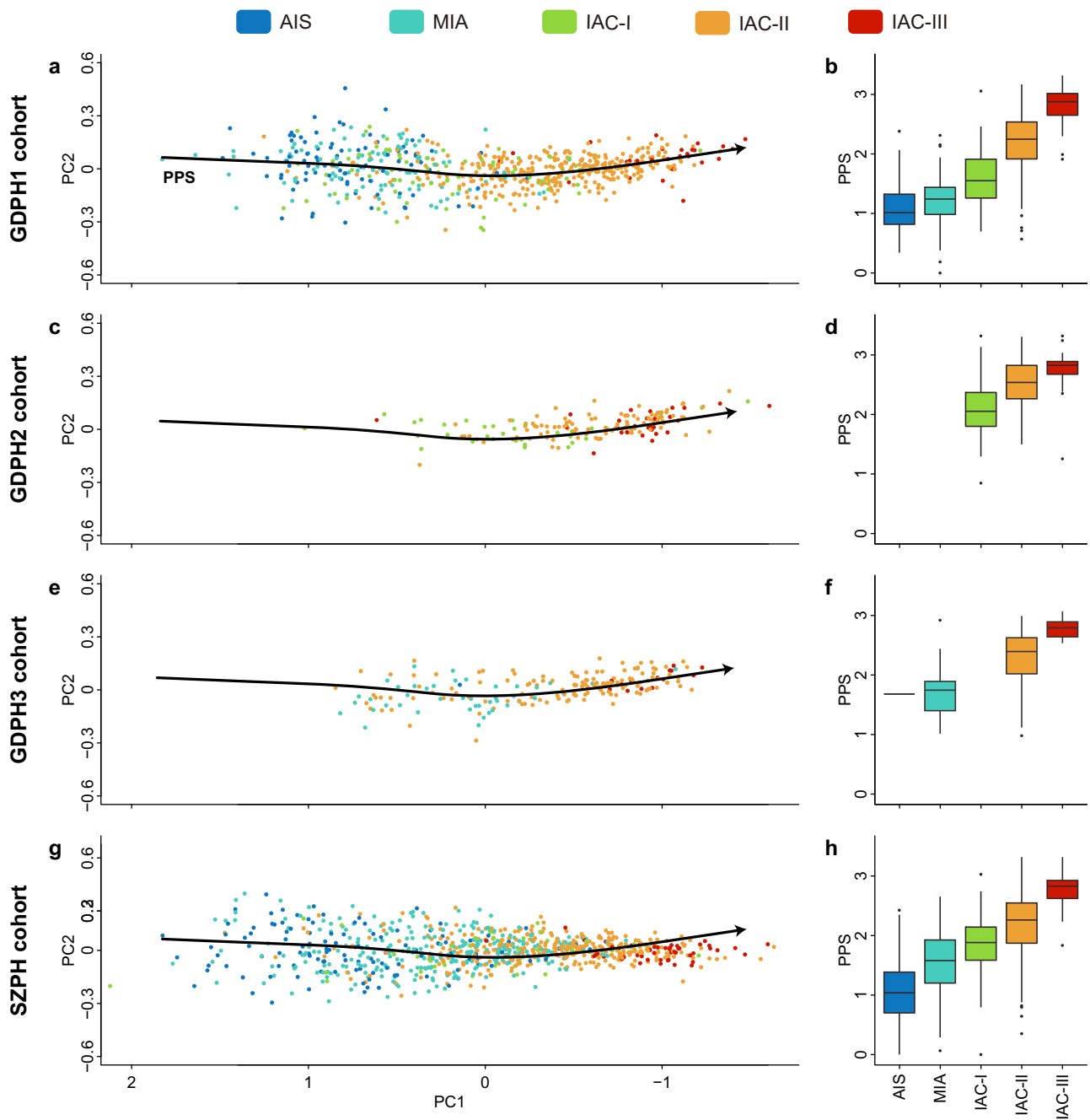


Fig. 2 | The radiomic trajectory of esLUAD progression and the distribution of corresponding pseudo-progression score (PPS). Left column: Tumor image embeddings of multi-institute patient cohorts in the PC space along the RadioTrace. Right column: Distribution of the inferred pseudo-progression score of each tumor.

The PPS is in accordance with the pathological grade records. **a, b** The GDPH1 cohort, in which the radiomic trajectory was constructed; **c, d** the GDPH2 cohort; **e, f** the GDPH3 cohort; **g, h** the SZPH cohort.

according to the median PPS value. We conducted similar radiogenomic analyses as illustrated above. Similar to the findings of tumors in all grades, the enriched pathways such as Glycolysis, Coagulation, EMT (Fig. S6, Tables S4 and S5) show that the two groups are at different stages of the tumor progression procedure. Next, we calculated the correlation between gene expression level and the PPS, and found that the highly-correlated genes, such as *SPP1*, *IGFBP3* and *AGER*, are genes important in esLUAD progression (Fig. S7). We also identified gene modules associated with RadioTrace within the IAC-II grade, and found a gene module “MEMidnightblue” that was associated with patient survival conditions (Fig. S8), while the other gene modules were not. The KEGG analysis also revealed that the genes

grouped within the PPS-correlated module are predominantly enriched in the Cell Cycle pathway (Table S6). All these results showed that RadioTrace is associated with genetic alterations and differential gene expressions related to esLUAD progression, providing a microscopic perspective for understanding the evolution of CT images during LUAD progression as well as the biological mechanism of RadioTrace.

RadioTrace helps track the dynamic progression of pulmonary nodules

The above results have demonstrated that CT images of esLUAD patients can be translated to RadioTrace, a radiomic trajectory which can help infer

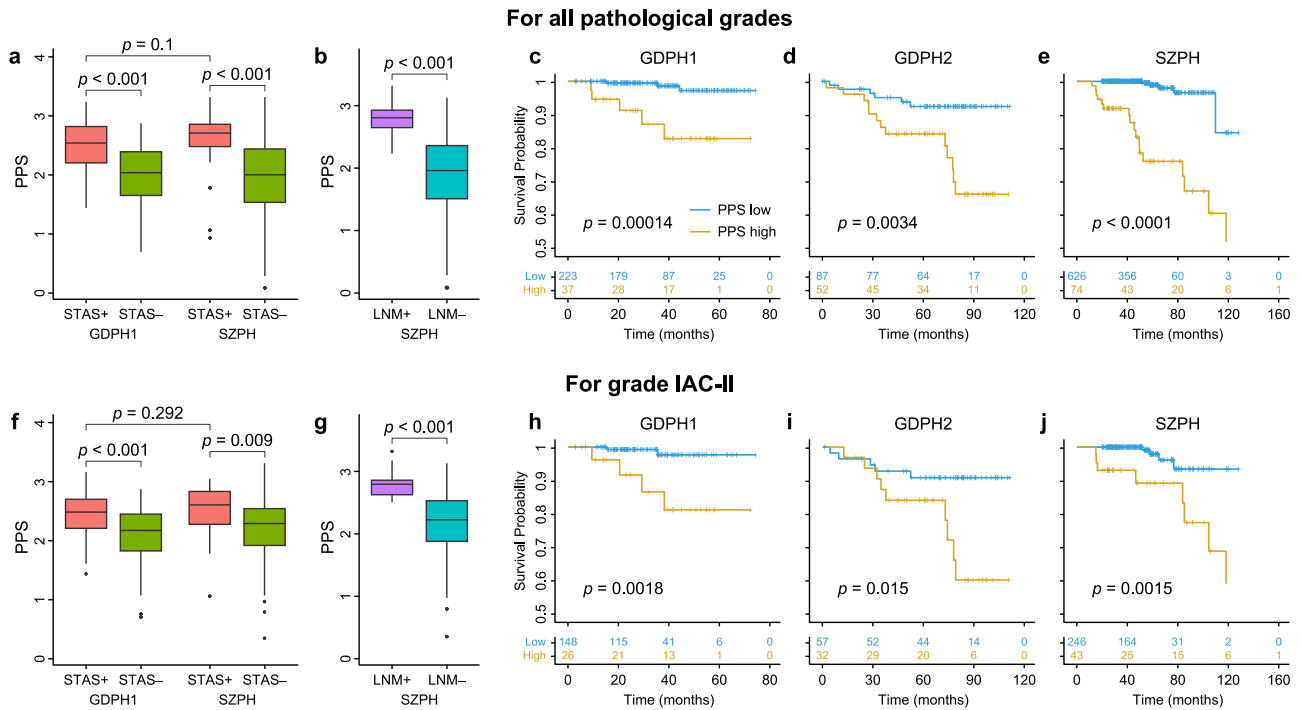


Fig. 3 | Association between PPS with tumor properties and patient prognosis in all esLUAD tumors and in grade IAC-II. PPS is significantly higher for tumors with spread through air space (STAS) and lymph-node metastasis (LNM). The disease-free survival (DFS) is significantly better for patients with lower PPS. **a–e** All esLUAD patients. **f–i** patients within grade IAC-II.

Table 2 | Univariate and multivariate analysis for DFS prediction using all patients in the combination of GDPH1 and GDPH2 cohorts

Variable	Univariate Analysis		Multivariate Analysis	
	HR (95% CI)	p-value	HR (95% CI)	p-value
Age	1.022 (0.983–1.061)	0.272		
Gender	2.132 (0.976–4.658)	0.058		
Smoking	1.120 (0.452–2.776)	0.807		
EGFR mutation	0.532 (0.241–1.175)	0.119		
Diameter(mm)	1.101 (1.032–1.175)	0.004	1.034 (0.960–1.114)	0.381
PPS	9.017 (3.115–26.10)	<0.001	7.254 (2.298–22.90)	0.001

Table 3 | Univariate and multivariate analysis for DFS prediction using all patients in the SZPH cohort

Variable	Univariate Analysis		Multivariate Analysis	
	HR (95% CI)	p-value	HR (95% CI)	p-value
Age	1.033 (0.997–1.071)	0.072		
Gender	2.757 (1.191–6.382)	0.018	2.229 (0.922–5.388)	0.075
Smoking	1.345 (0.394–4.593)	0.636		
EGFR mutation	0.660 (0.215–2.023)	0.467		
Diameter(mm)	1.141 (1.081–1.204)	<0.001	0.975 (0.898–1.059)	0.548
PPS	31.79 (9.482–106.6)	<0.001	37.78 (7.255–196.8)	<0.001

the progression of esLUAD. To further explore whether RadioTrace can assist in tracking the dynamic progression of pulmonary nodules for individualized monitoring and timely intervention, we retrospectively analyzed additional cases that were not included in the previously described cohorts.

Case 1 is a 55-year-old woman was initially found to have a ground-glass opacity nodule in the left upper lobe during a routine health

examination in 2018. Subsequently, she underwent CT scans regularly. Each CT scan of this nodule was projected to the embedding space of RadioTrace. As shown in Fig. 5a, the PPSs of these CT images gradually increased over time, indicating that the nodule was slowly progressing. In July 2023, the patient underwent segmentectomy. Postoperative pathologic examination indicated the nodule was grade-II LUAD composed of 70% acinar and 30%

Table 4 | Univariate and multivariate analysis for DFS prediction using IAC-II patients in the combination of GDPH1 and GDPH2 cohorts

Variable	Univariate Analysis		Multivariate Analysis	
	HR (95% CI)	p-value	HR (95% CI)	p-value
Age	1.033 (0.976–1.093)	0.268		
Gender	2.481 (0.746–8.247)	0.138		
Smoking	0.814 (0.178–3.719)	0.791		
EGFR mutation	0.334 (0.086–1.294)	0.112		
Diameter(mm)	1.087 (0.976–1.211)	0.131		
PPS	6.203 (1.304–29.52)	0.022	6.203 (1.304–29.52)	0.022

Table 5 | Univariate and multivariate analysis for DFS prediction using IAC-II patients in the SZPH cohort

Variable	Univariate Analysis		Multivariate Analysis	
	HR (95% CI)	p-value	HR (95% CI)	p-value
Age	0.988 (0.938–1.040)	0.644		
Gender	1.653 (0.553–4.941)	0.368		
Smoking	1.360 (0.292–6.337)	0.696		
EGFR mutation	0.698 (0.133–3.660)	0.671		
Diameter(mm)	1.009 (1.002–1.017)	0.017	1.002 (0.991–1.012)	0.771
PPS	11.89 (2.330–60.65)	0.003	9.553 (1.086–84.03)	0.042

lepidic growth patterns, without lymph node metastasis, pleural invasion or air space spread. A similar case with timely surgical intervention is detailed in the Supplementary Notes.

Case 2 is a 36-year-old man found to have a pure ground-glass opacity nodule in the right lower lung during a physical examination in August 2024. His physician recommended follow-up, and he underwent 6-monthly CT scans for one year. RadioTrace analysis showed a persistently low PPS with no significant change, indicating an indolent nodule at a very early stage (Fig. 5b). Later, the patient strongly requested surgery and ultimately underwent a segmentectomy in August 2025. Postoperative pathology confirmed a minimally invasive adenocarcinoma (MIA).

Case 3 is a 68-year-old woman found to have a pure ground-glass opacity nodule in the left upper lung during a physical examination in 2018 and was followed annually by CT. RadioTrace analysis showed that the PPS increased steadily over the years, with a marked rise in 2022, suggesting ongoing progression (Fig. 5c). However, surgery was not performed at that time. During later follow-ups, the PPS continued to increase, reaching its peak at the most recent examination. The patient eventually underwent surgery in August 2025, but intraoperative frozen pathology revealed mediastinal lymph node metastasis, necessitating lobectomy with systematic lymph node dissection instead of sublobar resection. Adjuvant treatment was still required after surgery.

Together, these cases illustrate distinct clinical scenarios. Case 1 represents an example of timely intervention that preserved organ function and achieved cure. In Case 2, continued surveillance might have been sufficient as a young man with a MIA had an extremely low risk of disease progression, but surgery was performed earlier than necessary. In Case 3, delayed surgery resulted in loss of an opportunity for curative resection. These findings suggest that RadioTrace has the potential to guide more precise management of pulmonary nodules, helping to avoid both over-treatment and undertreatment.

Discussion

In this study, we proposed a deep learning-based strategy to quantify esLUAD progression from CT images. We developed a contrastive learning framework to integrate radiomic and pathomic tumor information, and constructed a radiomic trajectory, RadioTrace, that transforms discrete pathological grades to the PPS, a continuous measurement of esLUAD

progression. Experiments on multi-institute patient cohorts showed that RadioTrace is in accordance with the order of esLUAD pathological grades and can uncover more detailed progression procedure within the same grade. RadioTrace is a significant indicator for tumor metastasis and patient’s survival conditions, and is correlated with the genetic and transcriptomic profiles related to tumor progression and prognosis. Furthermore, RadioTrace is highly consistent with the real progression of esLUAD. All the results suggested the effectiveness of RadioTrace and the corresponding PPS in quantifying esLUAD progression and their potential in clinical diagnosis and prognosis.

This work is a pilot study that quantifies tumor progression procedure in a continuous manner with radiomic approaches. Although some previous studies suggested that the dynamic evolution in the progression of LUAD is related to pathological histological images³⁸ or solid components in imaging³⁹, these manifestations at the pathologic level and at the image level may not align with each other¹⁵. There still lacks a quantitative and concordant indicator reflecting this evolution. As evident, RadioTrace was shown to be able to well predict phenotypes and prognoses of esLUAD patients, suggesting that it effectively captures pathological grade transitions and intra-grade differences. Furthermore, RadioTrace was shown to be associated with genetic alterations and differentially expressed genes, many of which are vital to esLUAD progression. All these findings demonstrated that RadioTrace may serve as such an effective radiomic indicator for quantifying the progression of esLUAD, consistent with trends at the pathological, genetic, and transcriptomic levels.

RadioTrace holds the potential to evolve into a valuable complementary tool for the clinical management of esLUAD. We found that tumors in AIS and MIA stages intertwined at the beginning of the radiomic trajectory, suggesting that they share similar features in CT images. Previous studies showed that there was no significant difference in the tumor mutational burden (TMB) and differentially expressed genes between pure ground-glass opacity-like (pGGO-like) and mixed-GGO-like (mGGO-like) AIS&MIA, suggesting that the impact of increased solid components on the genomic events of AIS&MIA is negligible⁴⁰. A genomic-level research also revealed an intertwined relationship between AIS and MIA²⁷. Clinically, Yotsukura et al. discovered a 100% disease-specific survival rate for AIS and MIA ten years’ post-surgery⁴¹. All these discoveries reflect the similar inert nature of AIS and MIA, which was captured by RadioTrace, suggesting that

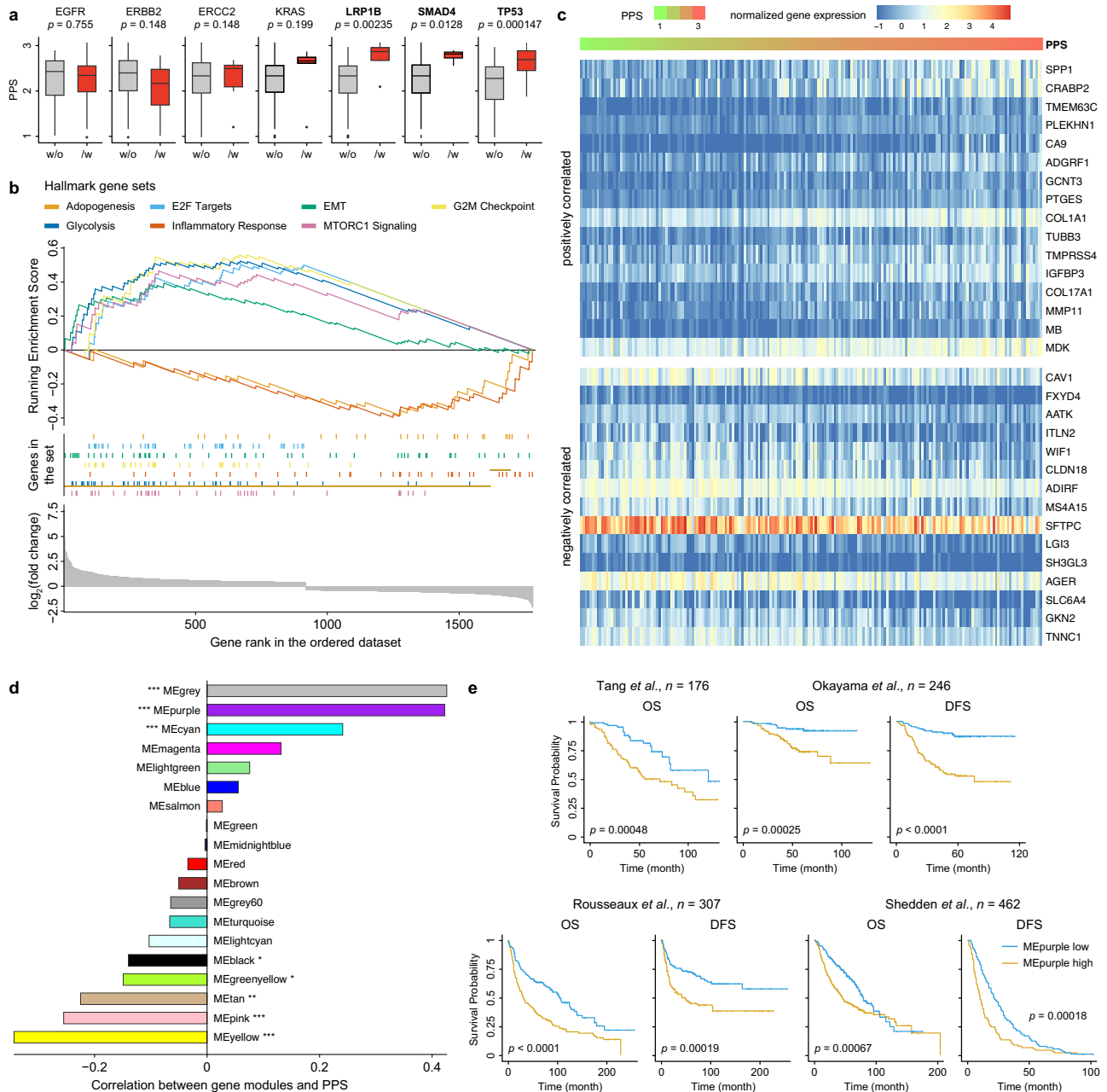


Fig. 4 | RadioTrace is associated with gene expression of esLUAD tumors. **a** Distribution of PPS between tumors with and without mutation of certain genes. **b** Enriched Hallmark gene sets identified by GSEA for tumors groups divided by the RadioTrace. **c** The correlation between gene expression and PPS of all tumors in the

GDPH3 cohort. **d** Correlation between WGCNA gene modules with PPS. **e** The identified gene module is associated with patient’s prognosis. * indicates a p -value less than 0.05, ** indicates a p -value less than 0.01, and *** indicates a p -value less than 0.001.

the homogeneous minimally invasive treatment modalities for these two pathological types of nodules is reasonable.

However, the increase in solid components proportion (consolidation tumor ratio, CTR) and the enlargement of diameter in invasive adenocarcinoma represent an elevation in the malignancy of the tumor⁴². Japan Clinical Oncology Group (JCOG) has conducted several randomized studies based on these indicators to develop corresponding diagnostic and therapeutic models^{43,44}. However, the definition of such indicators is ambiguous, especially in the absence of a unified measurement plane⁴⁵. This fuzzy definition leads to some controversy in using CTR for clinical decision-making. Ye *et al.* pointed out that CTR or solid component diameter cannot serve as prognostic factors for esLUAD, which presents as partially solid nodules⁴⁶. However, Koike *et al.* concluded that more than 75% CTR on CT was an independent recurrent factor⁴⁷. In light of these

challenges associated with traditional indicators such as CTR and solid component diameter, RadioTrace emerges as a promising solution to overcome the limitations in characterizing tumor malignancy. Its unique advantage lies in three-dimensional sampling, providing a more comprehensive understanding of nodules and uncovering intricate details that conventional measures might miss. As evidenced by transcriptome data, RadioTrace not only identifies pathways related to tumor occurrence covered by CTR but also reveals additional insights, such as the heightened activity of the EMT pathway in higher PPS tumors—information crucial for understanding the dynamic nature of tumor progression (Fig S5c). As we know, during EMT, the epithelial cells lose junction, reorganize their cytoskeletons and the reprogram their gene expression profiles to acquired motility and invasive phenotype⁴⁸. This highlights RadioTrace’s capacity to capture a more nuanced and detailed landscape of tumor development,

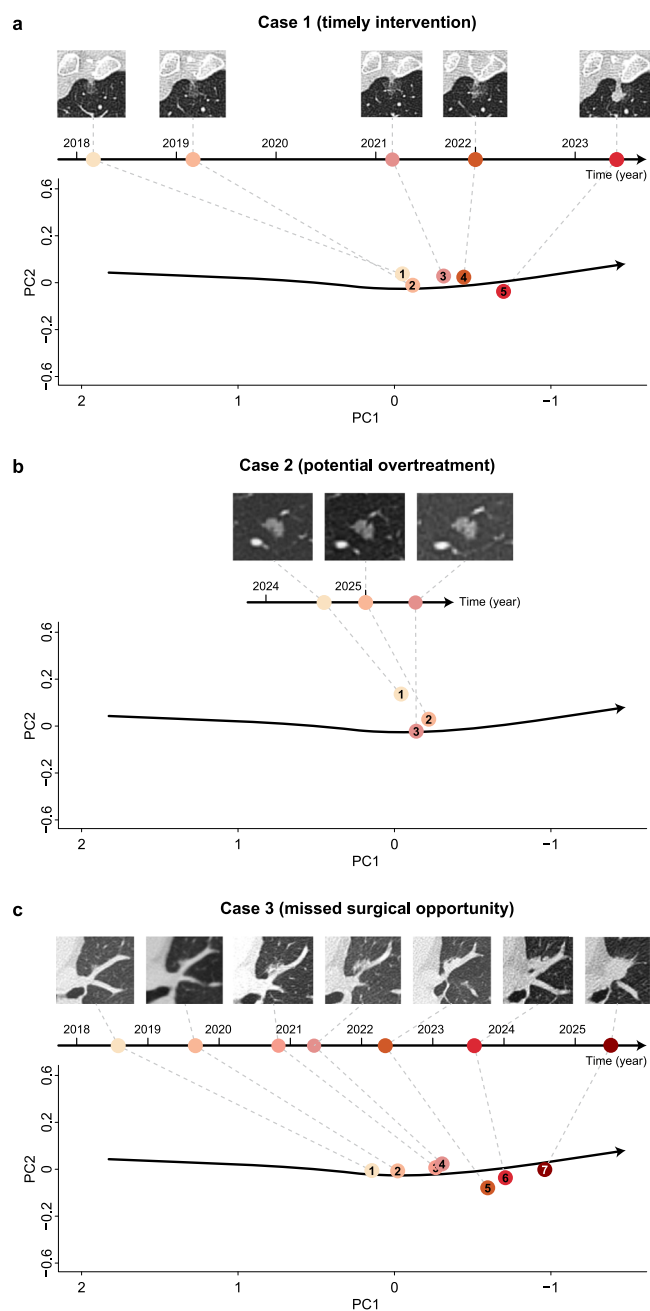


Fig. 5 | Longitudinal changes of individual tumors from three patients along with the RadioTrace. a Case 1 (timely intervention); b Case 2 (potential overtreatment); c Case 3 (missed surgical cured opportunity).

setting it apart as a valuable tool in enhancing diagnostic precision and therapeutic decision-making.

The outcomes of these cases underscore the potential value of RadioTrace in guiding the management of pulmonary nodules during follow-up. For early-stage lung cancer, surgical resection remains the only curative option. Yet, because ground-glass opacities (GGOs) often show indolent growth and favorable long-term survival, CT surveillance is commonly recommended for screen-detected subsolid nodules to balance oncologic benefit against surgical trauma and psychosocial burden^{49–52}. Current guidelines, however, primarily rely on tumor size and the extent of the solid component—criteria that are subject to interpretation and lack precise definition^{45,53}. As our cases illustrate, such conventional follow-up strategies may result in either premature surgery or missed opportunities for curative

intervention. By providing objective, quantitative measurements and capturing their temporal dynamics, RadioTrace offers a more reliable assessment of lesion status and reveals evolutionary patterns of nodules, thereby adding an important dimension to clinical decision-making.

Radiologic and pathologic examinations are two approaches routinely used in cancer diagnosis, providing the macroscopic and microscopic information, respectively. Integrating these cross-scale data modalities may benefit scientific studies and clinical decision-making. We employed the deep contrastive learning framework to integrate the radiomic and pathomic information. We used the pathological grades as labels and trained a CNN model to embed radiomic information into a latent space where the embeddings of tumors with the same label are as close as possible, and those with different labels are as far away as possible. This approach was shown to be effective and may be applied to radiopathomic studies of other cancer types. Moreover, pathological data such as H&E images may serve as a “mesoscopic” layer that bridges microscopic cellular features and macroscopic tissue-level texture of tumor, which is important to build more comprehensive cross-modality association of esLUAD and worth exploring in the future studies.

This study showed the capability of machine learning in discovering new knowledge from data. We did not provide the grade of tumors during model training, but the embeddings of tumors were correctly ordered in the PCA space according to pathological grades. Machine learning also assigned different spans on the radiomic trajectory for each grade, which helped us to discover the internal differences within IAC-II, which is a new insight for humans. During recent years, machine learning, especially deep learning, achieved great successes in many scenarios, such as computer vision and natural language processing⁵⁴. These successes most rely on existing knowledge used as supervision during model training. It is necessary to explore whether and to what extent machine learning can discover knowledge with minimum guidance or human intervention. Scientists have made some efforts in this direction and found that using machine learning only is able to uncover human embryonic cell development⁵⁵, heliocentric theory of solar system⁵⁶, etc. With the development of machine learning, more knowledge is expected to be discovered from data, and machine learning will become a powerful assistance for scientific research.

There are some important directions that remain to be explored in the future. The radiomic trajectory of later-stage LUAD progression can be further studied, which may be more complicated due to tumor heterogeneity. It is expected that uncovering later-stage tumor progression could help identify tumor subtypes and develop treatment plans. It is also worth analyzing the radiomic trajectories of other histological subtypes of lung cancer, such as squamous cell carcinoma, large cell carcinoma, etc., as well as their relationships with esLUAD progression. Also, with the accumulation of larger sample sizes, more powerful AI models could be used, and it is possible to build a radiomic–pathomic–genomic network to better illustrate the mechanism of tumor progression.

In addition, we observed the genomic and prognostic differences among patients of the IAC-II grade. More analyses with prospectively designed studies should be conducted to explore whether there are tumor subtypes within the same pathological grade, in which the histology images could be used to identify subtle morphological differences. Besides, although three retrospective long-term follow-up cases illustrated how RadioTrace may aid clinicians in managing pulmonary nodules, larger-scale clinical studies are still needed to define an optimal PPS cutoff for intervention and appropriate follow-up intervals, in order to balance the risks of overtreatment against those of delayed treatment.

As the clinical value of RadioTrace was evaluated on the Asian population in this study, international patient cohorts with larger sample sizes should be carried out to further validate and improve RadioTrace. Moreover, the radiomic features can be influenced by CT protocols, pre-processing procedures, and differences in tumor delineation, which may bias results. Therefore, it is important for future large-scale prospective studies not only to validate the clinical utility of RadioTrace in diverse

populations, but also to systematically assess these potential sources of variability and develop strategies to mitigate their impact, thereby improving robustness and generalizability.

Methods

Patient cohorts

In this study, we included CT images, clinical and pathological records of four esLUAD patient cohorts from Guangdong Provincial People's Hospital (GDPH) and Shenzhen People's Hospital (SZPH). As Fig. 1a illustrated, all included patients were experienced the quality control as the following inclusion criteria, (1) pathologically confirmed diagnosis of esLUAD, with pathological grades available; (2) pre-treatment CT images available; (3) patients received surgery alone. For each patient's CT images, the gross tumor volumes were delineated by an experienced radiologist using ITK-SNAP⁵⁷ and was confirmed by a senior radiologist. This study was approved with the ethics committee of GDPH (No. GDRHEC2019726H) and SZPH (No. KY-LL--2021584-01), and was conducted in accordance with ethical standards of the 1964 Helsinki Declaration and its later amendments. Patient's written informed consent was waived as no protected health information was used in this study.

Model of RadioTrace and calculation of PPSs

The acquisition of CT image and segmentation mask, as well as the pre-processing step were described in supplemental materials (*Supplementary notes*). The 3D tumor region was cropped from the original CT image using the segmentation mask, which was used as model input. The corresponding tumor grade was used as the label. We adopted the 3D version of ResNet-50⁵⁸ architecture, a classical CNN architecture, as the backbone for feature extraction from tumor images. Since the CT images were resized to a fixed voxel dimension during preprocessing, direct tumor size information was lost. To retain this information, we incorporated the measured 3D tumor volume from CT as an additional scalar input at the fully connected (FC) layers (Fig. 1). Specifically, we added another FC layer on the original ResNet-50 and concatenated the tumor volume with the output of the first FC layer in dimension 128. A ReLU activation layer and a dropout layer with a probability of 0.5 was added between two FC layers. The output embedding vector of the second FC layer with dimension 32 was used as the representation of each tumor.

As for model training, the triplet loss⁵⁹ was used as the loss function to minimize the Euclidean distances between positive pairs and maximize those between negative pairs (details in *Supplementary Notes*). The Adam optimizer was employed for parameter update with a learning rate of 0.001. We also adopted the learning rate decay strategy using the CosineAnnealingLR scheduler. The deep learning model, training and evaluation process was implemented with PyTorch (v1.12.1)⁶⁰ in Python 3.8, on a NVIDIA GeForce RTX 3090 GPU with 24GB of memory. The CUDA version is 11.4.

After model training, we obtained the embedding vectors of all the tumors in the GDPH1 cohort. We performed PCA and visualized the sample distribution using the first two PCs. We observed that tumors are located along a curve and the order of tumors is in accordance with their pathological grades. We employed the slingshot²³ method to infer the underlying trajectory on the first two PCs along which the tumors are distributed. The AIS cluster was set as the start and the slingshot method calculated the coordinates of the curve. The relative sample positions can be obtained by projecting them onto the trajectory. The quantitative progression measurement PPS of each sample was calculated by setting the PPS of the sample at the start of the trajectory as 0. In this way, a projection from CT image to the trajectory position, including the CNN model, the PCA projection, and the trajectory coordinate calculation, was constructed. For each sample in the test set, we calculated the corresponding embedding vector, the coordinate in the PC space and the position on the trajectory directly using the projection function.

Constructing trajectories with pre-defined radiomic features

A set of 110 radiomic features defined in pyradiomics⁶¹ were used, including 19 first-order features, 16 3D shape features and 75 texture features. The

detailed definition of each feature is available in the package document (<https://pyradiomics.readthedocs.io/en/latest/features.html>). For each tumor in the GDPH1 cohort, we extracted the 110 features from CT images. To reduce redundancy and mitigate multi-collinearity, we first applied low variance filtering (variance > 0.01), retaining 77 features. We then removed highly correlated features with pairwise correlation greater than 0.95, resulting in 43 features. These processed features were subsequently used for PCA, and we visualized the distribution of tumors and their pathological grades on the first two PCs.

DNA sequencing and processing

DNA was purified using the issue kit (Qiagen, Hilden, Germany) from the formalin fixed paraffin embedded (FFPE) samples following the manufacturer's instructions. Only samples harboring tumor cell contents above 20% were considered qualified and included. Fragments between 200–400 bp from DNA were purified (Agencourt AMPure XP Kit, Beckman Coulter, CA, USA), hybridized with capture probe baits, selected with magnetic beads, and amplified. Target capture was performed using commercial panels consisting of 520 genes (OncoScreen Plus) or 196 genes (Geneseq Prime). The samples were then sequenced, and the sequencing data were aligned to the hg19 human reference genome using Burrows-Wheeler Aligner (v0.7.10)⁶², Genome Analysis ToolKit (v3.2)⁶³ and VarScan (v2.4.3)⁶⁴ were used for local alignment optimization, duplication marking, and variant calling. Maftools⁶⁵ was used to analysis somatic variants. Overlapping genes between the two panels were used for analysis.

RNA sequencing and processing

Resected tumors were collected from December, 2021 to October, 2022, which were stored as frozen tissue and kept at -80°C in the tissue bank of GDPH. Total RNA was extracted from tumor tissue by Trizol reagent (Invitrogen) separately, and the RNA quality was checked by Bioanalyzer 2200 (Aligent) and kept at -80°C . The cDNA libraries were constructed for each pooled RNA sample using the NEBNext[®] Ultra[™] Directional RNA Library Prep Kit for Illumina according to the manufacturer's instructions. The products were purified and enriched by PCR to create the final cDNA libraries and quantified by Agilent2200. The tagged cDNA libraries were pooled in equal ratio and used for 150 bp paired-end sequencing in a single lane of the Illumina HiSeq X Ten. The experiments of Library construction and RNA Sequencing were completed at the center of Novellbio lab (Shanghai, China).

We processed RNA sequencing reads with the following steps. First, adapters of RNA-seq reads were trimmed using cutadapt (v 3.4)⁶⁶. Then, reads were aligned using STAR (v 2.5.3a)⁶⁷ with "--twopassMode Basic --outSAMstrandField introMotif --quantMode TranscriptomeSAM". The STAR index was built using the Gencode⁶⁸ annotation gtf file (v 32) and the corresponding genome reference (GRCh38.p13). We employed RSEM (v 1.3.0)⁶⁹ to quantify read counts with "--no-bam-output --paired-end". The output expression matrix was used for downstream analysis.

Statistical analysis

For correlation analysis, we calculated the Spearman correlation between two factors. We performed Wilcoxon rank-sum test to evaluate the difference between two groups for categorical variables.

For survival analysis, we employed Kaplan–Meier (K–M) plots with the log-rank test to visualize survival differences between two patient groups, and performed Cox regression to assess whether a factor was a significant predictor of patient prognosis.

The p -value was used as a quantitative indicator for the significance of difference, and p -value less than 0.05 was regarded as significant.

Differential expression and enrichment analysis

Given two groups of patients, we performed differential expression (DE) analysis using the edgeR package (v3.34.1)⁷⁰. The identified genes with adjusted p -value less than 0.05 were regarded as significant DE genes between two groups. Besides the DE genes identified by RadioTrace, we

performed DE analysis between tumors in invasive and non-invasive groups divided by pathological grades (AIS/MIA vs. IAC). We also downloaded the TCGA-LUAD⁷¹ dataset and calculated DE genes between tumors in N0 stage and the other N stages, as well as DE genes between M0 stage and the other M stages. We performed the Fisher's Exact Test (FET) to evaluate the overlap between DE genes identified by RadioTrace and genes in each functional gene set defined in the MSigDB.

The GO and KEGG enrichment analyses were implemented using the clusterProfiler package (v4.0.5)⁷². For GSEA, we ranked DE genes by their logarithmic fold changes and referred to the hallmark gene sets from Molecular Signatures Database (MSigDB)⁷³ to identify enriched gene sets. For GO and KEGG enrichment analysis, we selected DE genes with logarithmic fold changes larger 1.0 and smaller than -1.0 and identified functional gene sets with adjusted p -value less than 0.05.

Using Hallmark gene sets to predict the PPS with XGBoost

To further investigate the potential biological mechanisms underlying PPS, we performed single-sample gene set enrichment analysis (ssGSEA)⁷⁴ to calculate the enrichment scores of 50 Hallmark gene sets for each tumor sample. These enrichment scores were then used as input features to build an XGBoost regression model for predicting PPS. To interpret the contribution of each gene set to the model, we applied SHapley Additive exPlanations (SHAP) analysis and ranked the top 20 most important pathways.

Correlation analysis between PPS and gene expression

We first performed logarithm normalization for the gene expression values. Then we calculated the Spearman correlation between the normalized expression and the PPS for each gene. The corresponding p -value less than 0.05 was regarded as significant. We selected the top 15 positively and negatively correlated genes for illustration, respectively.

Gene modules construction and evaluation

We clustered genes with similar expression patterns into gene modules using the weighted graph co-expression network analysis (WGCNA) package (v1.71)³³. We chose the soft power beta according to the algorithm which gives the optimal fit for a scale-free topology. The soft power was set as 6. The gene modules were identified by the dynamic cutting tree algorithm implemented in this R package. Next, we calculated the Spearman correlation between the mean expression of each gene module and the PPS obtained from the radiomic trajectory. We chose the gene modules with correlation $r > 0.3$ and significance $p < 0.05$ as potential transcriptomic signature of the radiomic trajectory.

We evaluated the selected gene modules by testing whether they could stratify samples into subgroups with significantly different prognosis. We collected another four public LUAD patient cohorts with paired gene expression profile and survival records, including Okayama et al. (GSE31210)³⁴, Rousseaux et al. (GSE30219)³⁵, Tang et al. (GSE42127)³⁶ and Shedden et al. (GSE68465)³⁷. The association between selected gene modules and LUAD prognosis could further illustrate the effectiveness and generalization ability of the radiomic trajectory.

Data availability

The CT images and clinicopathological records data used in this study are available upon reasonable requests. The bulk RNA data that support the findings of this study are available from Omix (OMIX007086). Panel DNA-NGS data is available from Omix (OMIX007090).

Code availability

We have developed a Python package -- RadioTrace for users to quantify the progression procedure of esLUAD from medical images. The input of RadioTrace package is the CT image volume and corresponding segmentation mask of tumor, and the default output is a value indicating the progression status of the tumor using the constructed radiomic trajectory as a reference. The RadioTrace package supports input data format of either

dicom series (.dcm) or NIFTI (.nii), and tumor segmentation format of either RTStruct (.dcm) or NIFTI (.nii). Users can obtain intermediate results such as the embedding vector, and visualize the location of tumor on the radiomic trajectory and the relative position with other pathological-grade-labeled tumors. For more details, please check the webpage of this package: <https://pypi.org/project/RadioTrace>.

Received: 6 June 2025; Accepted: 23 September 2025;

Published online: 17 November 2025

References

- Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446–454 (2018).
- Clark, W. Tumour progression and the nature of cancer. *Br. J. Cancer* **64**, 631–644 (1991).
- The Lancet. Lung cancer: a global scourge. *Lancet* **382**, 659 (2013).
- Travis, W. D. et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma. *J. Thorac. Oncol.* **6**, 244–285 (2011).
- 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).
- 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).
- Haoran, E. et al. The IASLC grading system for invasive pulmonary adenocarcinoma: a potential prognosticator for patients receiving neoadjuvant therapy. *Ther. Adv. Med. Oncol.* **15**, 175883592211480 (2023).
- Mu, W. et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat. Commun.* **11**, 5228 (2020).
- Thunnissen, E. et al. Reproducibility of histopathological subtypes and invasion in pulmonary adenocarcinoma. An international interobserver study. *Mod. Pathol.* **25**, 1574–1583 (2012).
- Warth, A. et al. Interobserver variability in the application of the novel IASLC/ATS/ERS classification for pulmonary adenocarcinomas. *Eur. Respir. J.* **40**, 1221–1227 (2012).
- 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).
- Lami, K. et al. Standardized classification of lung adenocarcinoma subtypes and improvement of grading assessment through deep learning. *Am. J. Pathol.* **193**, 2066–2079 (2023).
- Park, B. J. et al. Proposal of a revised International Association for the Study of Lung Cancer grading system in pulmonary non-mucinous adenocarcinoma: The importance of the lepidic proportion. *Lung Cancer* **175**, 1–8 (2023).
- Pan, X. et al. The artificial intelligence-based model ANORAK improves histopathological grading of lung adenocarcinoma. *Nat. Cancer* **5**, 347–363 (2024).
- Qiu, Z.-B. et al. A novel radiopathological grading system to tailor recurrence risk for pathologic stage IA lung adenocarcinoma. *Semin. Thorac. Cardiovasc. Surg.* **35**, 594–602 (2023).
- Reck, M. & Rabe, K. F. Precision diagnosis and treatment for advanced non-small-cell lung cancer. *N. Engl. J. Med.* **377**, 849–861 (2017).
- Lee, G., Bak, S. H. & Lee, H. Y. CT radiomics in thoracic oncology: technique and clinical applications. *Nucl. Med. Mol. Imaging* **52**, 91–98 (2018).
- Cho, H., Lee, G., Lee, H. Y. & Park, H. Marginal radiomics features as imaging biomarkers for pathological invasion in lung adenocarcinoma. *Eur. Radio.* **30**, 2984–2994 (2020).

19. Xu, F. et al. Radiomic-based quantitative CT analysis of pure ground-glass nodules to predict the invasiveness of lung adenocarcinoma. *Front. Oncol.* **10**, 872 (2020).
20. Wu, L. et al. The value of various peritumoral radiomic features in differentiating the invasiveness of adenocarcinoma manifesting as ground-glass nodules. *Eur. Radio.* **31**, 9030–9037 (2021).
21. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
22. Liu, Y. et al. Image-based inference of tumor cell trajectories enables large-scale cancer progression analysis. *Sci. Adv.* **11**, eadv9466 (2025).
23. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* **19**, 477 (2018).
24. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
25. Gerstberger, S., Jiang, Q. & Ganesh, K. Metastasis. *Cell* **186**, 1564–1579 (2023).
26. Camp, R. L., Dolled-Filhart, M. & Rimm, D. L. X-Tile: a new bioinformatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin. Cancer Res.* **10**, 7252–7259 (2004).
27. Zhang, C. et al. Genomic landscape and immune microenvironment features of preinvasive and early invasive lung adenocarcinoma. *J. Thorac. Oncol.* **14**, 1912–1923 (2019).
28. Hu, C. et al. Genomic profiles and their associations with TMB, PD-L1 expression, and immune cell infiltration landscapes in synchronous multiple primary lung cancers. *J. Immunother. Cancer* **9**, e003773 (2021).
29. Wang, Y. et al. SMAD4 mutation correlates with poor prognosis in non-small cell lung cancer. *Lab. Investig.* **101**, 463–476 (2021).
30. Lei, R. et al. Potential role of PRKCSH in lung cancer: bioinformatics analysis and a case study of Nano ZnO. *Nanoscale* **14**, 4495–4510 (2022).
31. Kent, L. N. & Leone, G. The broken cycle: E2F dysfunction in cancer. *Nat. Rev. Cancer* **19**, 326–338 (2019).
32. Wang, Z. et al. Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. *Nat. Commun.* **12**, 6500 (2021).
33. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
34. Okayama, H. et al. Identification of genes upregulated in ALK -Positive and EGFR/KRAS/ALK -negative lung adenocarcinomas. *Cancer Res.* **72**, 100–111 (2012).
35. Rousseaux, S. et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* **5**, 186ra66–186ra66 (2013).
36. Tang, H. et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clin. Cancer Res.* **19**, 1577–1586 (2013).
37. Shedden, K. et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* **14**, 822–827 (2008).
38. Karasaki, T. et al. Evolutionary characterization of lung adenocarcinoma morphology in TRACERx. *Nat. Med.* **29**, 833–845 (2023).
39. Suzuki, K. et al. A prospective radiological study of thin-section computed tomography to predict pathological noninvasiveness in peripheral clinical IA lung cancer (Japan Clinical Oncology Group 0201). *J. Thorac. Oncol.* **6**, 751–756 (2011).
40. Shang, J. et al. Differences of molecular events driving pathological and radiological progression of lung adenocarcinoma. *eBioMedicine* **94**, 104728 (2023).
41. Yotsukura, M. et al. Long-term prognosis of patients with resected adenocarcinoma in situ and minimally invasive adenocarcinoma of the lung. *J. Thorac. Oncol.* **16**, 1312–1320 (2021).
42. Li, Y. et al. Genomic characterisation of pulmonary subsolid nodules: mutational landscape and radiological features. *Eur. Respir. J.* **55**, 1901409 (2020).
43. Saji, H. et al. Segmentectomy versus lobectomy in small-sized peripheral non-small-cell lung cancer (JCOG0802/WJOG4607L): a multicentre, open-label, phase 3, randomised, controlled, non-inferiority trial. *Lancet* **399**, 1607–1617 (2022).
44. Suzuki, K. et al. A single-arm study of sublobar resection for ground-glass opacity dominant peripheral lung cancer. *J. Thorac. Cardiovasc. Surg.* **163**, 289–301.e2 (2022).
45. Yoshiyasu, N., Kojima, F., Hayashi, K. & Bando, T. Radiomics technology for identifying early-stage lung adenocarcinomas suitable for sublobar resection. *J. Thorac. Cardiovasc. Surg.* **162**, 477–485.e1 (2021).
46. Ye, T. et al. Lung adenocarcinomas manifesting as radiological part-solid nodules define a special clinical subtype. *J. Thorac. Oncol.* **14**, 617–627 (2019).
47. Koike, T., Koike, T., Yamato, Y., Yoshiya, K. & Toyabe, S. Prognostic predictors in non-small cell lung cancer patients undergoing intentional segmentectomy. *Ann. Thorac. Surg.* **93**, 1788–1794 (2012).
48. Ang, H. L. et al. Mechanism of epithelial-mesenchymal transition in cancer and its regulation by natural compounds. *Med. Res. Rev.* **43**, 1141–1200 (2023).
49. Xing, X. et al. Decoding the multicellular ecosystem of lung adenocarcinoma manifested as pulmonary subsolid nodules by single-cell RNA sequencing. *Sci. Adv.* **7**, eabd9738 (2021).
50. Fu, F. et al. Distinct prognostic factors in patients with stage I non-small cell lung cancer with radiologic part-solid or solid lesions. *J. Thorac. Oncol.* **14**, 2133–2142 (2019).
51. Kobayashi, Y. How long should small lung lesions of ground-glass opacity be followed?. *J. Thorac. Oncol.* **8**, 309–314 (2013).
52. Barta, J. A. et al. The American Cancer Society National Lung Cancer Roundtable strategic plan: optimizing strategies for lung nodule evaluation and management. *Cancer* **130**, 4177–4187 (2024).
53. Chen, H. et al. The 2023 American Association for Thoracic Surgery (AATS) expert consensus document: management of subsolid lung nodules. *J. Thorac. Cardiovasc. Surg.* **168**, 631–647.e11 (2024).
54. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
55. Shah, N. et al. An experiment on Ab initio discovery of biological knowledge from scRNA-seq data using machine learning. *Patterns* **1**, 100071 (2020).
56. Iten, R., Metger, T., Wilming, H., del Rio, L. & Renner, R. Discovering physical concepts with neural networks. *Phys. Rev. Lett.* **124**, 010508 (2020).
57. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* **31**, 1116–1128 (2006).
58. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778, <https://doi.org/10.1109/CVPR.2016.90> (2016).
59. Hoffer, E. & Ailon, N. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition* (eds Feragen, A., Pelillo, M. & Loog, M.) 84–92 (Springer International Publishing, Cham, 2015).
60. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32 (eds Wallach, H. et al.) 32 8024–8035 (Curran Associates, Inc., Vancouver, British Columbia, Canada, 2019).
61. van Griethuysen, J. J. M. et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017).
62. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
63. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

64. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
 65. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
 66. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
 67. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
 68. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
 69. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
 70. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
 71. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
 72. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
 73. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
 74. Hänzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinforma.* **14**, 7 (2013).
- cohorts and provides the delineation of tumor area as well as the STAS and LNM diagnoses. J.L., Z.Q., and L.W. designed the methods. J.L. implemented the deep learning and the trajectory inference model. J.L., Q.M., L.W., and Z.Q. performed the bioinformatics analysis. J.L., Z.Q., S.D. and L.W. wrote the manuscript under the supervision of W.Z., X.Z., L.Y., and H.W.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-02031-0>.

Correspondence and requests for materials should be addressed to Lin Yang, Xuegong Zhang or Wen-Zhao Zhong.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 82241235 and 92470105), Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0529400), National Key R&D Program of China (2021YFF1200901), High-level Hospital Construction Project (DFJH201801), National High-Level Talents Special Support Program (KA0120231004), Guangdong Basic and Applied Basic Research Foundation (No. 2019B1515130002), Guangdong Provincial Key Laboratory of Lung Cancer Translational Medicine (2017B030314120), and Tsinghua-Fuzhou Institute of DataTechnology Project (JIDT2024022).

Author contributions

W.Z., X.Z., and L.Y. initiated and lead the project. Z.Q., S.D., M.W., H.L., C.Z., H.X., B.J., F.X., J.Y., Y.W., Y.L., and W.Z. collected the in-house patient