# Supplementary Materials

**Contents**

## 1. Supplementary Methods

### 1.1 Inclusion and exclusion criteria

We retrospectively reviewed data for 6860 patients in five academic medical centers. As Supplementary Figure 1 shown, this multi-center data came from different cities of China: Peking University People's Hospital is located in the north of China (Beijing city), and the other hospitals are located in the southwest of China (across three distinct cities in Sichuan Province). Supplementary Figure 2 shows the inclusion and exclusion criteria. In this study, 4332 patients with 19,300 CT scans underwent physical examination at West China Hospital of Sichuan University and Mianzhu People's Hospital were incorporated as the cohort 1 for self-supervised pre-training. 2028 non-small-cell lung cancer (NSCLC) patients with follow-up data and underwent CT examination before initial treatment at Peking University People's Hospital and Chengdu Shangjin Nanfu Hospital were enrolled in the cohort 2 for multi-task learning network fine-tuning and internal validation. 500 NSCLC patients with follow-up data and underwent CT examination before initial treatment at Guangan People's Hospital were incorporated as cohort 3 for external validation. We followed,[1] to exclude the cases with unqualified CT images or missing values of clinical data. Specifically, in the cohort 1 for self-supervised pre-training, we excluded 1323 unqualified CT images, therefore obtained 17,977 CT scans. In the cohort 2 for the development of multi-task learning network, we excluded 19 patients with missing values of clinical data and 121 patients with unqualified CTs, therefore obtained 1177 patients for fine-tuning, 711 patients for internal validation. In the cohort 3 for external validation of our method, we excluded 8 patients with missing values of clinical data and 24 patients with unqualified CTs, therefore obtained an external validation set including 468 patients.

### 1.2 CT-based self-supervised pre-training framework using contrastive learning

To learn general visual representations from 3D CT images, we built a self-supervised learning CTCLR (CT-based Contrastive Learning of Representations) framework, which was pre-trained on large amounts of unlabeled CT images. Through contrastive learning, CTCLR learned the visual representation by distinguish positive CT pairs against negative ones. The proposed CTCLR consists of four steps, including 3D image augmentation, 3D visual feature extraction, non-linear projection and contrastive learning.

Given a 3D CT image $x_i$ from a minibatch of $N$ samples, the image is firstly transformed into two correlated views $\widetilde{x_i}$ and $\widetilde{x_j}$ by using the 3D image augmentation strategies. We employed six 3D augmentation strategies including random Gaussian noise, random Gaussian blur, random brightness transform, random contrast transform, random Gamma transform and random crop. The augmented views from the same CT image are clustered together, whereas others are pushed further away.

Then we employed 3D ResNet,[2] as the feature extractor $f(\cdot)$ to map the CT images into vector representations and multi-layer perceptron (MLP) with one hidden layer followed by ReLU activation function as the non-linear projection head $g(\cdot)$ to project feature representations into the space where contrastive loss was applied.

$$z_i = g(f(\widetilde{x_i}))$$

Finally, the normalized temperature-scaled cross entropy loss (NT-Xent),[3] is regarded as the contrastive loss applied into $2N$ data points to maximize the similarity of positive pairs. The loss function is defined as

$$l_{(i,j)} = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} exp(sim(z_i, z_k)/\tau)}$$

where $N$ is the batch size and $\tau$ denotes the temperature parameter of which value is set to 0.1. $z_i$ and $z_j$ are the vector representations from a positive pair generated by the non-linear projection head. $1_{[k\neq i]}$ is the indicator function of which value is set to 1 when $k \neq i$. We followed SimCLR,[3] to implement cosine similarity as $sim(z_i, z_j) = \frac{{z_i}^T z_j}{\|z_i\|\|z_j\|}$.

**1.3 Multi-task learning network**

We developed a multi-task learning network to predict a patient's OS risk, the probability of early stage (stage I), and EGFR genotype. The proposed multi-task learning network had three branches, including one primary branch for OS risk prediction and two extra branches for auxiliary tasks (early stage and EGFR genotype detection). We employed ResNet,[2] as the backbone of our multi-task learning network. Given a 3D CT image $x_i$, the proposed multi-task learning network $h(\cdot)$ generated three predicted value, including OS risk $s_i$, the probability of EGFR mutation $m_i$ and the probability of early stage $p_i$.

$$s_i, m_i, p_i = h(x_i; \theta)$$

where $\theta$ represents the parameter of the network $h(\cdot)$.

The rationale for combining these tasks in a unified model is that EGFR genotype and cancer stage have

strong association with patients' prognosis. We hypothesized that integrated these three tasks into a unified network might improve the prediction accuracy of OS risk. Thus, we designed a fusion layer to share feature representations obtained from different tasks in different branches at multi-scale levels. Specifically, the input was a 3D tensor concatenated from three feature maps obtained from corresponding network branches. Then the concatenated 3D tensor was input into two convolution layers. Both layers use a $3\times3$ kernel and followed by a ReLU activation function. By adopting fusion layers, different feature representations of prognosis-related tasks were shared and fused, which might provide primary task with more useful prognostic knowledge.

For the prediction of OS risk, we leveraged the negative log partial likelihood,[4] as the loss function to optimize the primary branch of the multi-task learning network:

$$L_s = -\sum_{i:E_i=1}\left(s_i - log\sum_{j\in\Re(T_i)} e^{s_i}\right)$$

where the values $T_i$, $E_i$, $s_i$ are the respective event time, event indicator and the predicted death rate for the $i^{th}$ observation. $\{i: E_i = 1\}$ represents the set of patients with an observation event of death ($E_i = 1$). $\Re(T_i) = \{i: T_i \geq t\}$ is the set where patients are still alive at time $t$.

For the prediction of the other two auxiliary tasks, cross entropy loss was employed as the loss function:

$$L_{aux_1} = L_{ce}(m_i, Y_i^m)$$

$$L_{aux_2} = L_{ce}(p_i, Y_i^p)$$

where $L_{aux_1}$ and $L_{aux_2}$ represent the loss function for the EGFR genotype detection and early stage prediction tasks respectively. $Y_i^m$ and $Y_i^p$ are the corresponding labels of the two auxiliary tasks.

Therefore, the objective loss $\mathcal{L}$ to optimize the entire multi-task learning network is the following:

$$\mathcal{L} = \alpha L_s + \beta L_{aux_1} + \gamma L_{aux_2}$$

where $\alpha$, $\beta$ and $\gamma$ are the soft parameters controlling the loss of three different tasks. To give the higher importance to the primary task of OS risk prediction, we used $\alpha = 0.4$, $\beta = 0.3$ and $\gamma = 0.3$.

### 1.4 Training details

In this study, the proposed IPES was implemented on Pytorch and trained on NVIDIA A10 GPUs. For the self-supervised pre-training framework, CTCLR was updated by AdamW optimizer with a learning rate of $10^{-3}$, a weight decay of $10^{-5}$ and a batch size of 1,024. The number of maximum epochs was set

to 200. For the down-stream multi-task learning-based fine-tuning, model was updated by Adam optimizer with a learning rate of $3\times10^{-3}$, a batch size of 64 and the fine-tuning epochs of 50.

### 1.5 Sample size estimation and study protocol

According to the previous study,[5] which predicts the OS for stage I-IIIA NSCLC, the 5-year death rate of the subgroup (A) to receive additional survival benefit is 39.1%, whereas subgroup (B) showed a higher 5-year death rate of 83.3%. The ratio of the number of patients in subgroup (A) to subgroup (B) is approximately 5.75. Based on the above discussion, the results of the estimated sample sizes by statsmodels (Version: 0.14.0) on Python (Version: 3.9.1) are 43 and 57 for the two subgroups (power = 0.90), with a two-sided alpha of 0.05.

This study corresponds to the previously registered study (ChiCTR1800015700), and it was a continued study based on the above-registered study. Therefore, this study was not additionally registered as a clinical trial.

### 1.6 Software materials

In this study, we exploited several software and packages for data analysis and visualization: (1) Our IPES and Grad-CAM algorithm was implemented using PyTorch (Version: 1.7.1) on Python (3.9.1). (2) The cutoff IPES-score was selected by X-tile software (Version: 3.6.1). (3) We developed the Cox proportional hazard model by using lifelines package (Version: 0.27.0) on Python (Version: 3.9.1). (4) Sample size evaluation was performed using statsmodels (Version: 0.14.0) on Python (Version: 3.9.1).

## 2. Supplementary References


1 Deng KX, Wang L, Liu YC et al. A deep learning-based system for survival benefit prediction of tyrosine kinase inhibitors and immune checkpoint inhibitors in stage IV non-small cell lung cancer patients: A mutlicenter, prognostic study. *EClinicalMedicine* 2023; **51**: 101541.

2 He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016; 770-778.

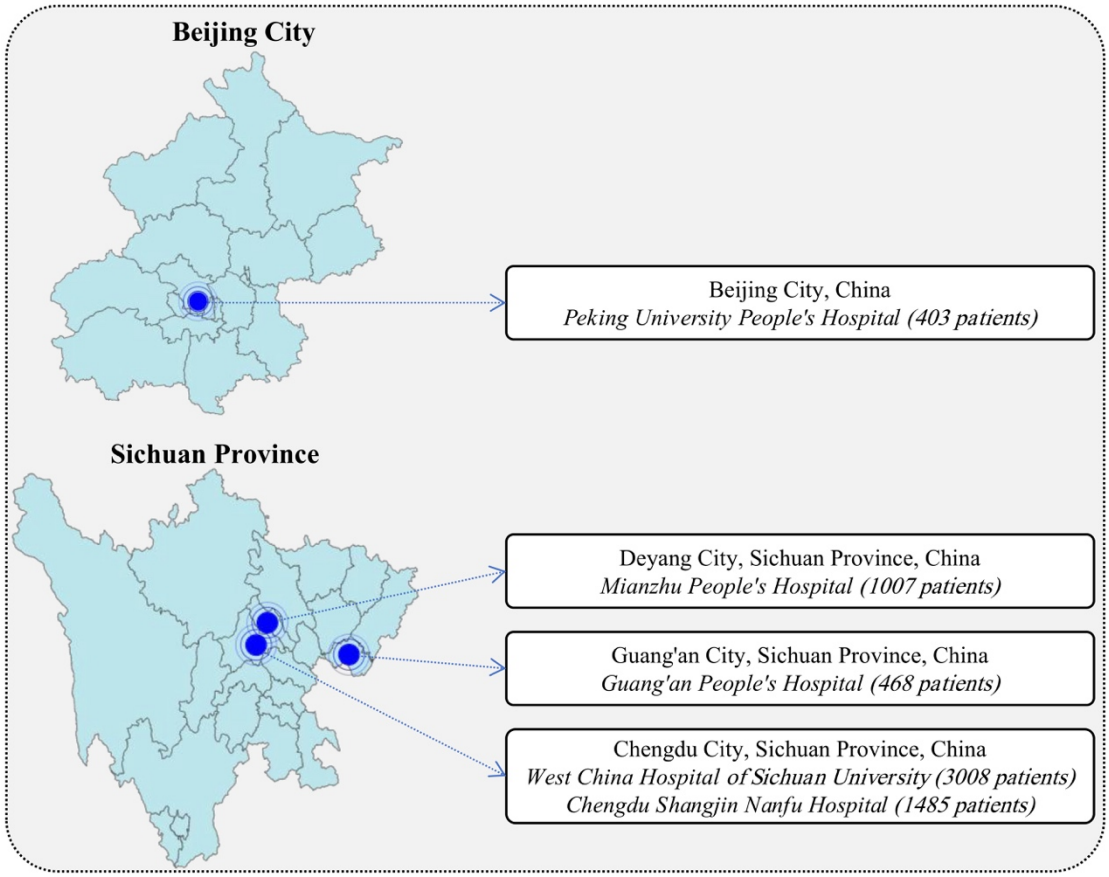
3 Chen T, Kornblith, S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. *In 2020 International Conference on Machine Learning (PMLR)* 2020; 1597-1607.

4 Katzman JL, Shaham U, Cloninger A, Bates J, Jiang TT, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology* 2018; **18**: 24.

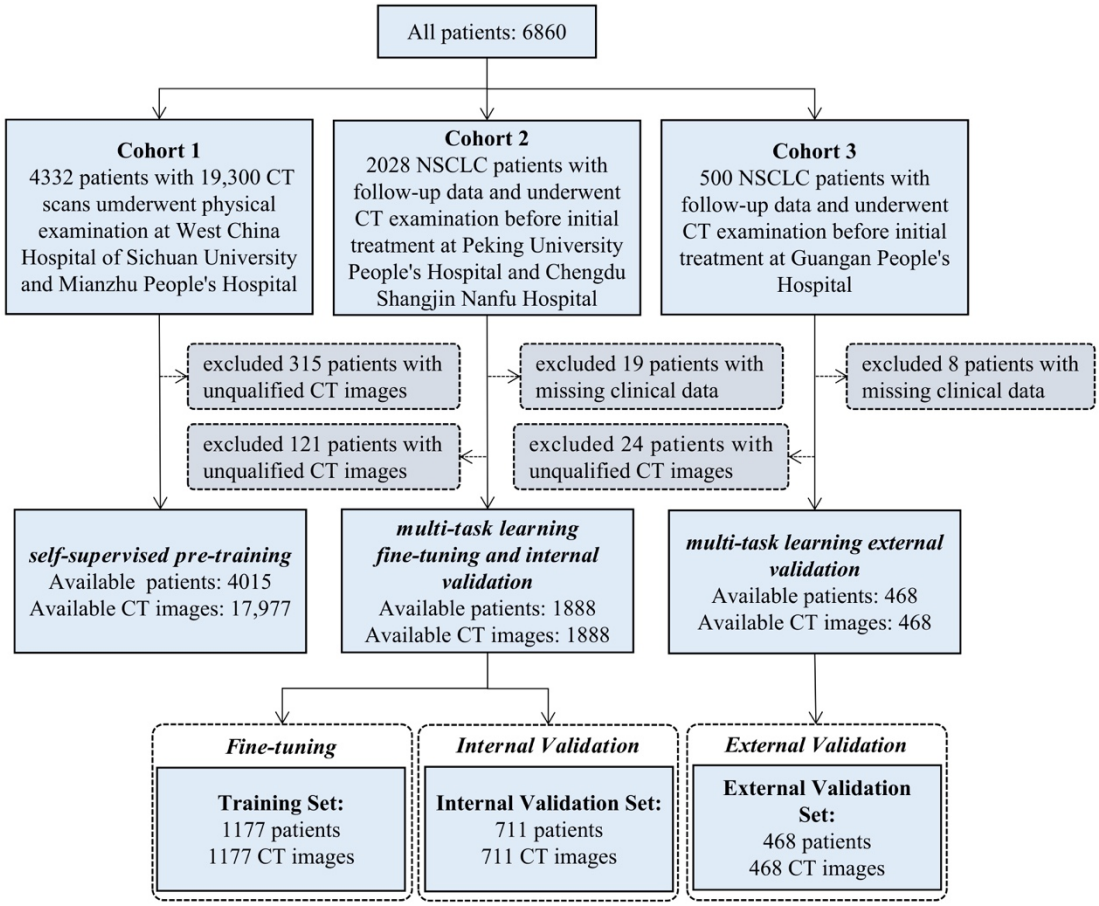
5 Zheng SY, Guo JP, Langendijk JA et al. Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning. *Radiotherapy and Oncology* 2023; **180**: 109483.

## 3. Supplementary Figures

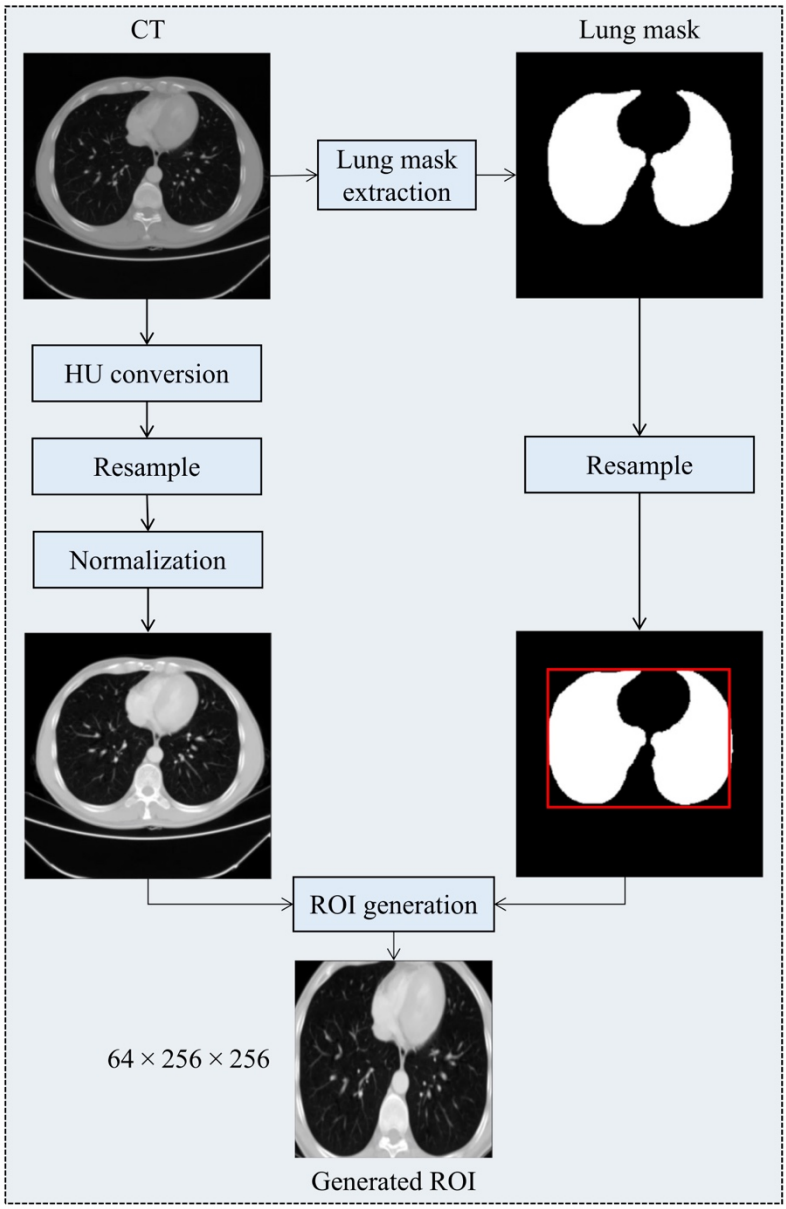
**Supplementary Figure S1. Maps describing the hospitals' location along with the corresponding number of patients in this study.**



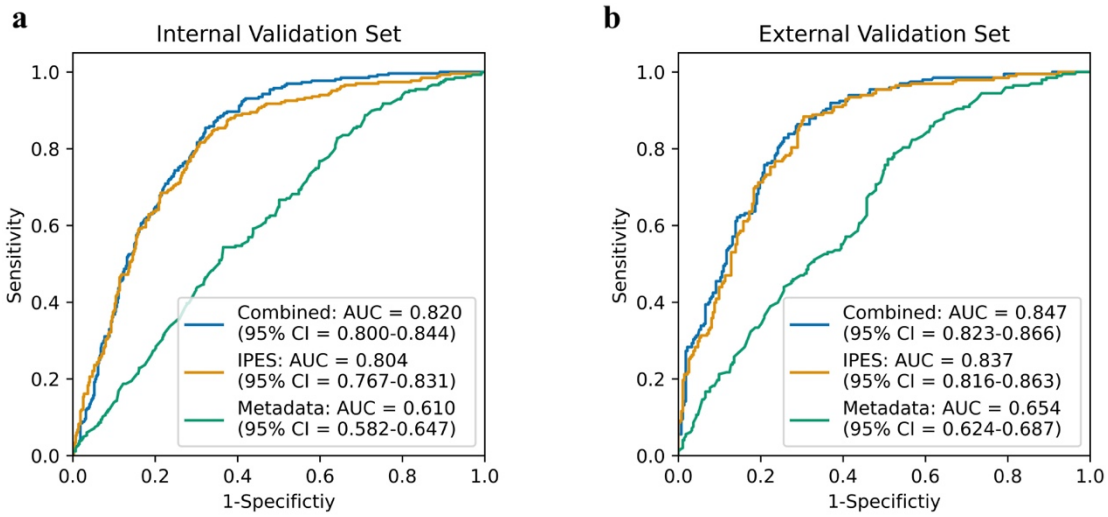
**Supplementary Figure S2. Flowchart escribing the datasets included in this study for self-supervised pre-training and multi-task learning.** Patients inclusion and exclusion criteria were also considered.



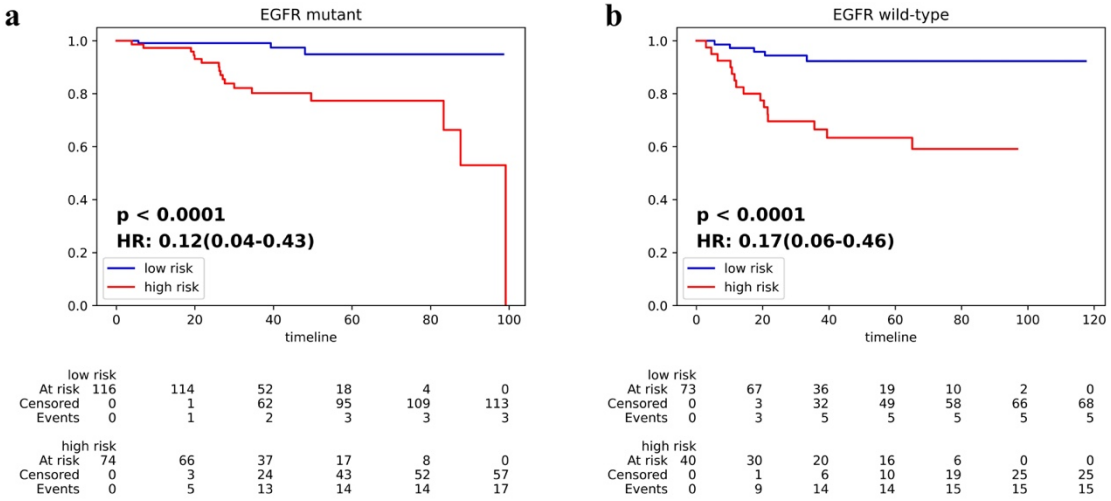
**Supplementary Figure S3. Pre-processing procedure of CT images.** Our CT image pre-processing procedure consists of five major steps including lung mask extraction, HU conversion, image resampling, intensity normalization and ROI generation.



**Supplementary Figure S4. IPES performance on detecting early-stage.** (a)-(b) ROC curves represent the performance of the IPES, metadata-based model and the combined model in the (a) internal validation set and (b) external validation set.



**Supplementary Figure S5. IPES performance on predicting OS risk of stage I-III resected NSCLC patients with *EGFR* mutation or *EGFR* wild-type status.** (a)-(b) Kaplan-Meier curves of the low- and high-risk subgroups predicted by IPES for stage I-III resected NSCLC patients with (a) *EGFR* mutation and (b) *EGFR* wild-type status in the external validation set. In the Kaplan-Meier curves, the horizontal axis represents survival time (months) and the vertical axis represents survival probability.




| | | | | | | |
|---|---|---|---|---|---|---|
| low risk | | | | | | |
| At risk | 116 | 114 | 52 | 18 | 4 | 0 |
| Censored | 0 | 1 | 62 | 95 | 109 | 113 |
| Events | 0 | 1 | 2 | 3 | 3 | 3 |
| high risk | | | | | | |
| At risk | 74 | 66 | 37 | 17 | 8 | 0 |
| Censored | 0 | 3 | 24 | 43 | 52 | 57 |
| Events | 0 | 5 | 13 | 14 | 14 | 17 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| low risk | | | | | | | |
| At risk | 73 | 67 | 36 | 19 | 10 | 2 | 0 |
| Censored | 0 | 3 | 32 | 49 | 58 | 66 | 68 |
| Events | 0 | 3 | 5 | 5 | 5 | 5 | 5 |
| high risk | | | | | | | |
| At risk | 40 | 30 | 20 | 16 | 6 | 0 | 0 |
| Censored | 0 | 1 | 6 | 10 | 19 | 25 | 25 |
| Events | 0 | 9 | 14 | 14 | 15 | 15 | 15 |

## 4. Supplementary Tables

**Supplementary Table S1. Stage I-III resected patients' characteristics in the training, internal validation and external validation sets.**

| | Training set (n=705) | Internal validation set (n=433) | External validation set (n=303) |
|---|---|---|---|
| **Sex** | | | |
| **Male** | 308 (43.6%) | 192 (44.3%) | 130 (42.9%) |
| **Female** | 397 (56.4%) | 241 (55.7%) | 173 (57.1%) |
| **Age, years** | 58 (48–68) | 59 (49–69) | 58 (49–67) |
| **Smoking status** | | | |
| **Former** | 222 (31.4%) | 137 (31.6%) | 101 (33.3%) |
| **Never** | 483 (68.6%) | 296 (68.4%) | 202 (66.7%) |
| **Cancer family history** | | | |
| **Yes** | 49 (6.9%) | 42 (9.6%) | 23 (7.5%) |
| **No** | 656 (93.1%) | 391 (90.4%) | 280 (92.5%) |
| **Tumor family history** | | | |
| **Yes** | 83 (11.7%) | 52 (12.0%) | 33 (10.8%) |
| **No** | 622 (88.3%) | 381 (88.0%) | 270 (89.2%) |
| **Histology** | | | |
| **Adenocarcinoma** | 636 (90.2%) | 391 (90.3%) | 273 (90.0%) |
| **Others** | 69 (9.8%) | 42 (9.7%) | 248 (10.0%) |
| **Stage** | | | |
| **I** | 433 (61.4%) | 258 (59.6%) | 190 (62.7%) |
| **II-III** | 272 (38.6%) | 175 (40.4%) | 113 (37.3%) |
| **EGFR genotype** | | | |
| **mutant** | 442 (62.6%) | 296 (68.3%) | 190 (62.7%) |
| **wild-type** | 263 (37.4%) | 137 (31.7%) | 113 (37.3%) |
| **Death status** | | | |
| **Dead** | 89 (12.6%) | 61 (14.0%) | 40 (13.2%) |
| **Censored** | 616 (87.4%) | 372 (86.0%) | 263 (86.8%) |

Data are n (%) or mean (SD). EGFR= epidermal growth factor receptor. Cancer family history = family history of lung cancer. Tumor family history = family history of other cancers (excluding lung cancer).

**Supplementary Table S2. C-index of the IPES on predicting OS risk for stage I-III resected NSCLC patients in the training, internal validation and external validation sets.**

| | C-index (95% CI) |
|---|---|
| **Training set** | 0.806 (0.744–0.846) |
| **Internal validation set** | 0.783 (0.744–0.825) |
| **External validation set** | 0.817 (0.786–0.849) |

C-index=concordance index. CI=confidence interval.

**Supplementary Table S3. C-index comparison of the IPES (combine TNM stage with MTL-score) with the TNM staging system (using stage alone) on predicting OS risk for stage I-III resected NSCLC patients.** The internal validation set was employed to compare the IPES with another baseline.

| | C-index (95% CI) |
|---|---|
| **IPES** | 0.783 (0.744–0.825) |
| **TNM staging system** | 0.733 (0.684–0.788) |

IPES=intelligent prognosis evaluation system. C-index=concordance index. CI=confidence interval.

**Supplementary Table S4. C-index comparison of the IPES (based on self-supervised pre-training and multi-task learning) with the multi-task learning method (without self-supervised learning) and single-task learning method (without both self-supervised pre-training and multi-task learning) on predicting OS risk for stage I-III resected NSCLC patients.** The internal validation set was employed to compare the IPES with other baselines.

| | C-index (95% CI) |
|---|---|
| **IPES** | 0.783 (0.744–0.825) |
| **Multi-task learning** | 0.765 (0.720–0.816) |
| **Single-task learning** | 0.741 (0.695–0.798) |

IPES=intelligent prognosis evaluation system. C-index=concordance index. CI=confidence interval.

**Supplementary Table S5. Performance comparison on early-stage prediction based on the IPES, metadata-based model and combined model in the training, internal validation and external validation sets.**

| Method | AUC (95% CI) | Accuracy (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|
| **Training Set** | | | | |
| **Combined** | 0.901 (0.883–0.918) | 0.808 (0.786–0.828) | 0.777 (0.744–0.809) | 0.860 (0.829–0.892) |
| **IPES** | 0.846 (0.822–0.870) | 0.765(0.732–0.792) | 0.769 (0.736–0.797) | 0.760 (0.710–0.801) |
| **Metadata-based** | 0.781 (0.751–0.806) | 0.701 (0.675–0.726) | 0.688 (0.653–0.724) | 0.722 (0.678–0.762) |
| **Internal Validation Set** | | | | |
| **Combined** | 0.820 (0.800–0.844) | 0.742 (0.712–0.771) | 0.698 (0.666–0.731) | 0.816 (0.783–0.854) |
| **IPES** | 0.804 (0.767–0.831) | 0.739 (0.715–0.762) | 0.694 (0.660–0.720) | 0.813 (0.769–0.853) |
| **Metadata-based** | 0.610 (0.582–0.647) | 0.600 (0.575–0.633) | 0.635 (0.600–0.665) | 0.543 (0.497–0.585) |
| **External Validation Set** | | | | |
| **Combined** | 0.847 (0.823–0.866) | 0.779 (0.750–0.800) | 0.744 (0.708–0.782) | 0.828 (0.796–0.857) |
| **IPES** | 0.837 (0.816–0.863) | 0.771 (0.745–0.798) | 0.707 (0.668–0.740) | 0.859 (0.823–0.893) |
| **Metadata-based** | 0.654 (0.624–0.687) | 0.607 (0.583–0.645) | 0.498 (0.455–0.540) | 0.758 (0.718–0.792) |

IPES=intelligent prognosis evaluation system. AUC=area under the curve. CI=confidence interval.

**Supplementary Table S6. Death rate for the low- and high-risk subgroups predicted by IPES on stage I and II-III resected NSCLC patients in the internal validation set.**

| | Number of participants | Number of events (5-year) | Number of events | Death rate (5-year) | Death rate | Hazard ratio (95% CI) | p-value |
|---|---|---|---|---|---|---|---|
| **Stage I** | | | | | | | |
| **low risk** | 230 | 8 | 8 | 3.5% | 3.5% | 0.10 (0.04–0.26) | 0.0029 |
| **high risk** | 28 | 9 | 10 | 32.1% | 35.7% | 4.25 (1.96–9.19) | NA |
| **overall** | 258 | 17 | 18 | 6.6% | 7.0% | NA | NA |
| **Stage II-III** | | | | | | | |
| **low risk** | 94 | 14 | 15 | 14.9% | 16.0% | 0.44 (0.24–0.83) | 0.012 |
| **high risk** | 81 | 26 | 28 | 32.1% | 34.6% | 2.26 (1.20–4.25) | NA |
| **overall** | 175 | 40 | 43 | 22.9% | 24.6% | NA | NA |

NA=not applicable. CI=confidence interval.

**Supplementary Table S7. Performance comparison on *EGFR* genotype prediction based on the IPES, metadata-based model and combined model in the training, internal validation and external validation sets.**

| | AUC (95% CI) | Accuracy (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|
| **Training Set** | | | | |
| **Combined** | 0.917 (0.903–0.933) | 0.825 (0.794–0.843) | 0.864 (0.827–0.895) | 0.802 (0.773–0.831) |
| **IPES** | 0.844 (0.817–0.864) | 0.748 (0.725–0.774) | 0.776 (0.740–0.807) | 0.731 (0.696–0.756) |
| **Metadata** | 0.777 (0.750–0.805) | 0.729 (0.704–0.753) | 0.711 (0.673–0.752) | 0.740 (0.706–0.766) |
| **Internal Validation Set** | | | | |
| **Combined** | 0.851(0.823–0.874) | 0.763 (0.740–0.787) | 0.714 (0.662–0.753) | 0.790 (0.760–0.820) |
| **IPES** | 0.819 (0.793–0.840) | 0.726 (0.703–0.744) | 0.777 (0.733–0.828) | 0.698 (0.666–0.732) |
| **Metadata** | 0.739 (0.711–0.764) | 0.691 (0.669–0.720) | 0.671 (0.625–0.726) | 0.702 (0.675–0.733) |
| **External Validation Set** | | | | |
| **Combined** | 0.824 (0.799–0.844) | 0.735 (0.708–0.761) | 0.725 (0.688–0.765) | 0.741 (0.697–0.776) |
| **IPES** | 0.791 (0.761–0.825) | 0.722 (0.697–0.745) | 0.791 (0.749–0.828) | 0.678 (0.643–0.709) |
| **Metadata** | 0.710 (0.677–0.745) | 0.682 (0.659–0.710) | 0.593 (0.549–0.634) | 0.737 (0.702–0.776) |

IPES=intelligent prognosis evaluation system. AUC=area under the curve. CI=confidence interval.

**Supplementary Table S8. C-index of the IPES on predicting OS risk for stage I-III resected NSCLC patients with *EGFR* mutation and *EGFR* wild-type status.**

| | ***EGFR* mutation (C-index (95% CI))** | ***EGFR* wild-type (C-index (95% CI))** |
|---|---|---|
| **Training set** | 0.824 (0.777–0.868) | 0.760 (0.714–0.809) |
| **Internal validation set** | 0.778 (0.723–0.824) | 0.762 (0.726–0.802) |
| **External validation set** | 0.803 (0.774–0.837) | 0.793 (0.768–0.827) |

*EGFR*= epidermal growth factor receptor. C-index=concordance index. CI=confidence interval.

**Supplementary Table S9. Death rate for the low- and high-risk subgroups predicted by IPES on stage I and II-III resected NSCLC patients with *EGFR* mutation status in internal validation set.**

| | Number of participants | Number of events (5-year) | Number of events | Death rate (5-year) | Death rate | Hazard ratio (95% CI) | p-value |
|---|---|---|---|---|---|---|---|
| **Stage I** | | | | | | | |
| **low risk** | 183 | 5 | 5 | 2.7% | 2.7% | 0.09 (0.03–0.30) | <0.0001 |
| **high risk** | 14 | 5 | 5 | 35.7% | 35.7% | 5.09 (1.74–14.90) | NA |
| **overall** | 197 | 10 | 10 | 5.1% | 5.1% | NA | NA |
| **Stage II-III** | | | | | | | |
| **low risk** | 89 | 16 | 18 | 18.0% | 20.2% | 0.26 (0.10–0.67) | 0.0045 |
| **high risk** | 10 | 5 | 6 | 50.0% | 60.0% | 3.92 (1.53–10.02) | NA |
| **overall** | 99 | 21 | 24 | 21.2% | 24.2% | NA | NA |

NA=not applicable. CI=confidence interval.

**Supplementary Table S10. Death rate for the low- and high-risk subgroups predicted by IPES on stage I and II-III resected NSCLC patients with *EGFR* wild-type status in internal validation set.**

| | Number of participants | Number of events (5-year) | Number of events | Death rate (5-year) | Death rate | Hazard ratio (95% CI) | p-value |
|---|---|---|---|---|---|---|---|
| **Stage I** | | | | | | | |
| **low risk** | 51 | 3 | 3 | 5.9% | 5.9% | 0.13 (0.03–0.53) | 0.0042 |
| **high risk** | 10 | 4 | 5 | 40.0% | 50.0% | 8.54 (1.97–37.00) | NA |
| **overall** | 61 | 7 | 8 | 11.5% | 13.1% | NA | NA |
| **Stage II-III** | | | | | | | |
| **low risk** | 52 | 7 | 7 | 13.5% | 13.5% | 0.23 (0.09–0.58) | 0.00081 |
| **high risk** | 24 | 12 | 12 | 50.0% | 50.0% | 4.71 (1.79–12.42) | NA |
| **overall** | 76 | 19 | 19 | 25.0% | 25.0% | NA | NA |

NA=not applicable. CI=confidence interval.