Early-Stage NSCLC Patients' Prognostic Prediction with Multi-information Using Transformer and Graph Neural Network Model

Authorship: MSc Jie Lian[1†], MD Jiajun Deng[2†], Dr Sai Kam Hui[3], Dr Mohamad Koohi-Moghadam[4], Dr Yunlang She[2], Dr Chang Chen[2*], Dr Varut Vardhanabhuti[1*]

[1]Department of Diagnostic Radiology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

[2]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China

[3]Department of Rehabilitation Science, Faculty of Health and Social Science, The Hong Kong Polytechnic University, China

[4]Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China

***Corresponding Authors:**

Varut Vardhanabhuti, MBBS BSc, FRCR, PhD

Clinical Assistant Professor, Department of Diagnostic Radiology, HKU

Mailing address: Room 406, Block K, Queen Mary Hospital, Pokfulam Road, Hong Kong SAR, China

Phone number: (+852) 2255 3307

Email: varv@hku.hk

Chang Chen, FACS, MD, PhD

Professor, Department of Thoracic Surgery, Tongji University School of Medicine

Mailing address: 507 Zhengmin Road, Shanghai, China

Phone number: (+86) 021-65115006

Email: changchenc@tongji.edu.cn

[†]These authors contributed equally to this work.

Abstract

Background: We proposed a population graph with Transformer-generated and clinical features for the purpose of predicting overall survival and recurrence-free survival for patients with early-stage NSCLC and to compare this model with traditional models.

Methods: The study included 1705 patients with lung cancer (stage I and II), and a public dataset for external validation (n=127). We proposed a graph with edges representing non-imaging patient characteristics and nodes representing imaging tumour region characteristics generated by a pretrained Vision Transformer. The model was compared with a TNM model and a ResNet-Graph model. To evaluate the models' performance, the area under the receiver operator characteristic curve (ROC-AUC) was calculated for both overall survival (OS) and recurrence-free survival (RFS) prediction. The Kaplan–Meier method was used to generate prognostic and survival estimates for low- and high-risk groups, along with net reclassification improvement (NRI), integrated discrimination improvement (IDI), and decision curve analysis (DCA). An additional subanalysis was conducted to examine the relationship between clinical data and imaging features associated with risk prediction.

Results: Our model achieved AUC values of 0·785 (95 % CI:0·716 - 0·855) and 0·695 (95 % CI:0·603 - 0·787) on the testing and external datasets for OS prediction, and 0·726 (95 % CI:0·653 - 0·800) and 0·700 (95 % CI:0·615 - 0·785) for RFS prediction. Additional survival analyses indicated that our model outperformed the present TNM and ResNet-Graph models in terms of net benefit for survival prediction.

Conclusion: Our Transformer-Graph model was effective at predicting survival in patients with early-stage lung cancer, which was constructed using both imaging and non-imaging clinical features. Some high-risk patients were distinguishable by using a similarity score function defined by non-imaging characteristics such as age, gender, histology type, and tumour location, while Transformer-generated features demonstrated additional benefits for patients whose non-imaging characteristics were non-discriminatory for survival outcomes.

Funding: There was no funding source for this study.

Keywords: Lung cancer, Graph convolutional networks, Vision Transformer, Survival Prediction

Introduction

Lung cancer is expected to account for more than 1·80 million deaths worldwide in 2021, making it the top cause of cancer-related mortality [1]. In early-stage (stage I and II) non-small cell lung carcinomas (NSCLC), surgical resection remains the therapy of choice. However, almost 40% to 55% of these tumours recur following surgery [2]. The clinical care of lung cancer patients would substantially benefit from accurate prognostic evaluation. Currently, TNM staging system of lung cancer based on the anatomic extent of disease is well recognised and widely adopted, which allows tumours of comparable anatomic extent to be grouped together [3]. Staging guides treatment and provides a broad prediction of prognosis, however individual characteristics, histology, and/or therapy characteristics may impact survival results, as seen by variation within stage groups. In the refinement of the staging system, non-anatomical predictors such as gene mutations and biomarker profiles were proposed to be incorporated [4]. However, the gene profiling approach relies on tissue sampling, and in addition, may not fully explain the intratumoural heterogeneity seen in NSCLC. Besides, such tests have barriers in deploying to routine oncology workflows due to high turnaround time, complexity, and cost [5].

To predict the patient's prognosis and to optimise individual clinical management, prognostic predictors such as TNM system and imaging-based high throughput quantitative biomarkers, radiomics, have been widely used to describe tumours [6-12]. Artificial Intelligence (AI) methods, especially some deep learning (DL) models, have recently been regarded as potentially valuable tools [13-15]. DL models generated multiple quantitative assessments for tumour characteristics, which have the potential to describe tumour phenotypes with more predictive power than the clinical model [15]. While the anatomical structures in a medical image are functionally and mechanically related, most AI-based methods do not take these interdependencies and relationships into account. This leads to instability and poor generalisation of performance [16]. With recent advancements in AI technology, several novel models have been proposed. Notably, the Transformer [17] model permits exceptional capabilities in natural language processing fields such as language translation and was later applied to the computer vision field and outperformed all state-of-the-art models given large amounts of training data [18]. This provides an intuitive reason to apply the Transformer model to the medical image to generate additional meaning for tumour features, as images were processed in sequence with inherent interdependencies [19].

The majority of current prognostic prediction methods have focused mainly either specific to their own domains, such as focusing solely on imaging data, whereas in clinical practice non-imaging clinical data such as sex, age, and disease history all play critical roles in disease prognosis prediction [20]. Although some researchers have used multi-modal techniques [21] to combine that information, it is not easy to explain how the various types of data interacted and how they contributed to the final prediction. Due to their lack of explanatory power, those models may not be easily applied in clinical practice [22]. Another type of neural network, called a graph neural network (GNN) [23], which deals with data that has a graph structure, enables researchers to create more flexible ways to embed various types of data. For example, nodes and edges in a graph might represent a variety of different types of data (imaging and clinical demographics information), and analysing these entities reveals the role of various data sources.

In this study, we proposed a GNN-based model that leverages imaging and non-imaging data for the prediction of the survival of patients with early-stage NSCLC. Patients were represented as a population graph, whereby each patient

corresponded to a graph node and was associated with a tumour feature vector that was learnt from the Transformer model, and graph edge weights between patients were derived from a similarity score that was derived from phenotypic data, such as demographics, tumour location, cancer type and TNM staging. This population graph was used to train a GraphSAGE [24] model for classifying individual patient's risk of overall survival and recurrence. Additionally, we attempted to determine the relative importance of imaging and non-imaging features within this model. The proposed model was trained and tested on a large dataset, followed by external validation using a publicly available dataset.

Methods

Participants

The study included consecutive patients who received surgery for early-stage NSCLC between January 2011 and December 2013 who matched the criteria. Inclusion criteria included: (1) pathologically proven stage I or stage II NSCLC; (2) preoperative thin-section CT image data; and (3) complete follow-up survival data. Patients undergoing neoadjuvant therapy were excluded from the study. The study protocol was approved by the Shanghai Pulmonary Hospital's Institutional Review Board and informed consent was waived owing to retrospective nature. Additionally, patients who met our criteria were retrieved from the NSCLC Radiogenomics [25] dataset as an external validation set (see Supplementary Figure 1 for the internal and external inclusion criteria flowchart).

We only used patients' initial CT scans in this study. For the main cohort, all CT scans were acquired using Somatom Definition AS+ (Siemens Medical Systems, Germany) and iCT256 (Siemens Medical Systems, Germany) (Philips Medical Systems, Netherlands). All image data were rebuilt using a 1 mm slice thickness and a 512×512 mm matrix. Intravenous contrast was administered in accordance with institutional clinical practice. Clinical data in this study were manually collected from medical records and were anonymised. Outpatient records and telephone interviews were used to collect follow-up data. The period between the date of surgery and the date of death or the final follow-up was defined as overall survival (OS). Recurrence-free survival (RFS) was calculated from the date of surgery to the date of recurrence, death, or last follow-up. (More details about internal scan parameters and follow-up strategies can be found in Supplementary II).

Image Annotation and Pre-processing

Patients' tumour region was manually labelled by experienced radiologists using 3D Slicer [26], with a centre seed point defining a bounding box. The regions of interest (ROIs) were first annotated by two junior thoracic surgeons (Y.S. and J.D. with 5 and 3 years of experience, respectively), then the consensus on ROI was obtained by a discussion with a senior radiologist (with more than 25 years of experience).

For image pre-processing, we first normalised all CT images and removed the surrounding noises such as bones by manual thresholding. The size of all tumour segments was fixed to 128mm × 128mm × 64mm. Small tumours were zero-padded. To reduce the computational cost, we resized the padded segments into 64mm × 64mm × 36mm and subsequently resized them as 2D square images (each row contained 6 tumour slices) with the size of 384mm × 384mm as shown in Figure 1A.

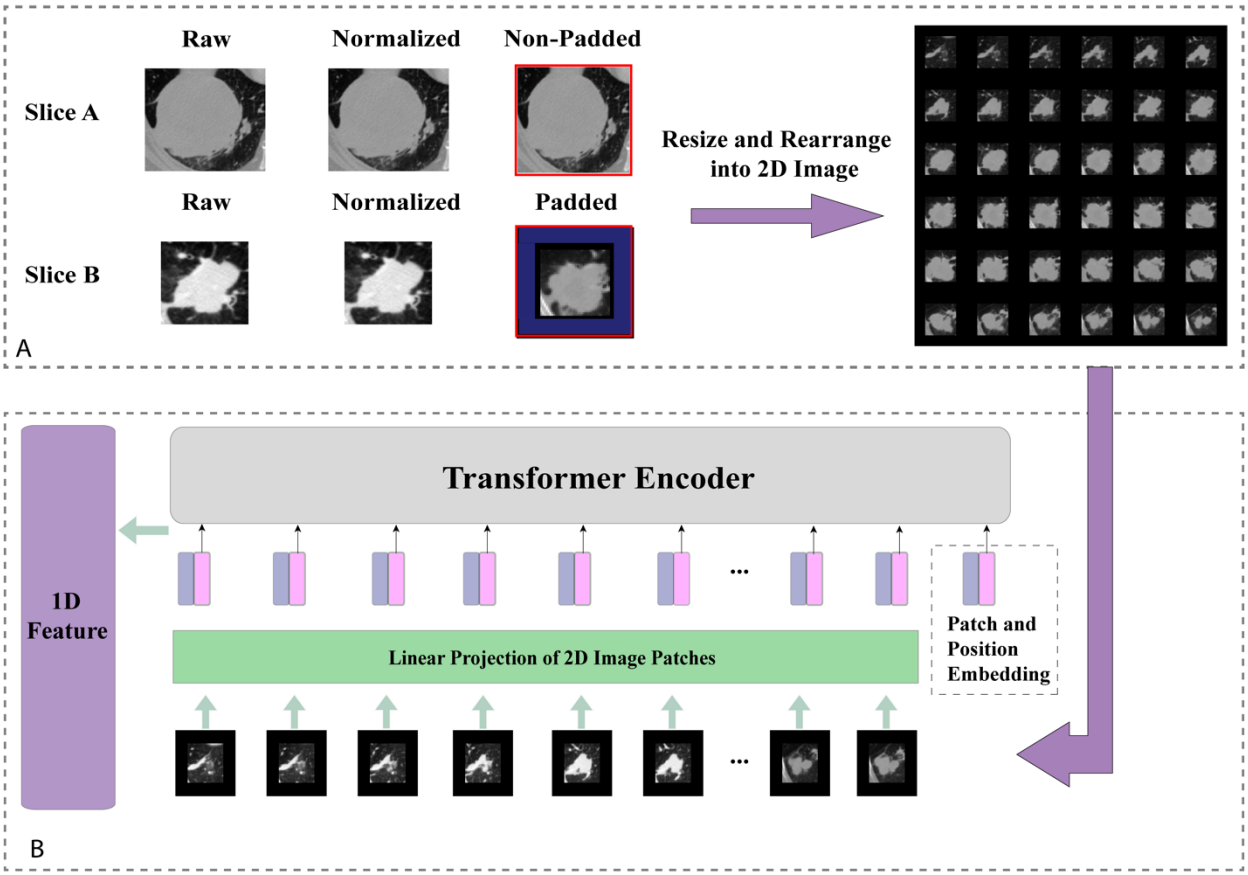


Figure 1. Tumour image processing and feature generation. (A) Tumour images normalization, reshaping and padding to standard sizes, then re-arranged into 2D images, (B) Generating 1D Transformer survival features from pretrained Transformer model.

Tumour Transformer Feature Generator

When pretrained on a large dataset and transferred to image recognition benchmarks, it has been shown that Vision Transformer (ViT) can achieve excellent results while requiring significantly less computational resources to train than state-of-the-art convolutional models [18]. To this end, we reasoned that by replacing the traditional CNN feature generator architecture with a Transformer structure could be an approach to produce meaningful survival-relevant features. In this study, we used a ViT pretrained on a large-scale dataset (ImageNet-21k [27]) as the feature generator, which takes 2D tumour segments as inputs. To meet the standard requirements of the sequence model, the input images were divided into 36 ordered patches and position embedding in the first step, followed by a linear projection function before entering the Transformer Encoder. We replaced the original classification layer with a fully connected layer to generate a 1D feature vector. The detailed implementation is illustrated in Figure 1B. The 1D feature vector was then assigned as the node feature for the individual patient in the graph network.

Patient survival graph network

A population graph method was used to leverage imaging and non-imaging data. Each patient was regarded as a node in a graph and its edge with neighbour was derived from a similarity score which was determined by the product between 4 component scores, namely demographics (gender and age), tumour location, cancer type (histology) and TNM staging (For more detail, refer to the supplementary for a detailed explanation of similarity scores). Two patients would be connected to each other if they shared similar component scores. The features of an individual patient (node feature) were obtained from the Transformer Encoder trained on the tumour images mentioned above.

Graph-based Neural Network Structure

We applied a graph-based deep neural network structure called GraphSAGE in this study. The proposed network took the whole population graph, along with the edge and node features as the input and generated a risk score in the last layer for each patient node as the output (see Figure 2). Within the network, every node feature was updated by an aggregation of information from its neighbours and itself, while the importance of different neighbours varied by the corresponding edges' weight.

We applied a two-layer GraphSAGE and global meaning pooling structure, aiming to allow each patient's information to be updated, first from its second neighbours and then its neighbours and itself consequentially. In order to emphasise the target of survival prediction, we specifically replaced the cross-entropy loss with Cox proportional hazards loss function [28] which both considered the survival time and events when training the network. The proposed network was implemented in Python, using the Deep Graph Library (DGL) with Pytorch backend.

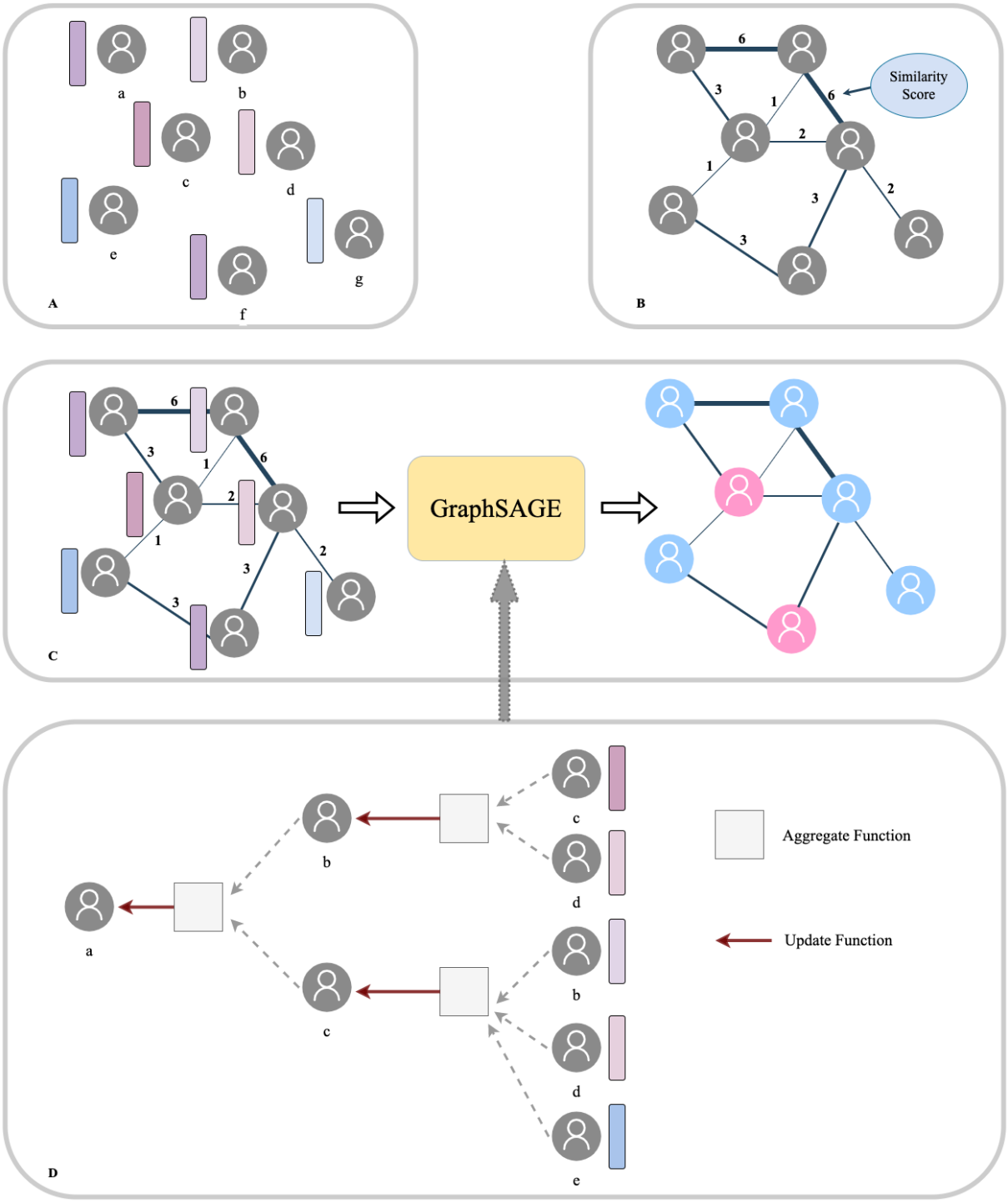


Figure 2. Population graph building and model prediction pipeline. (A) Each patient was regarded as a node and the Transformer-generated feature was regarded as node features. (B) Graph edges and the relevant weights were defined by their Similarity scores. (C) We then put the whole population graph to train the GraphSAGE network in order to make a prediction for each patient (pink indicates high risk and blue indicates low risk). (D) Node updating inside the GraphSAGE network.

Statistics Analysis

All patients from the main dataset were randomly separated into training, validation and testing sets with the proportion of 75%, 12·5% and 12·5% separately. We also tested the model on the external validation dataset. The proposed model was compared with the TNM staging system which was generally used in clinical practice and a ResNet-Graph model which has the same graph structure as our proposed model while the node feature was generated by a pretrained ResNet-18 model [29,30]. Some code

To evaluate whether there were statistically significant variations in survival between positive and negative groups, the area under the receiver operator characteristic curve (AUC) was determined for OS and RFS prediction to compare the models' performance. The Kaplan–Meier (KM) method was used to generate prognostic and survival estimates for groups with low and high risk (both for OS and RFS), which were stratified according to the training set's median prediction probability, with the log-rank test employed to establish statistical significance. To quantify the net benefits of survival prediction, we quantified the net reclassification improvement (NRI) and integrated discrimination improvement (IDI), as well as performed a decision curve analysis. All of the analyses above were performed in Python using the Lifelines package.

An additional subanalysis was performed on the test dataset to explore the relationship between patients' clinical information and imaging features contributing to risk prediction. We generated a sub-graph visualisation using PyVis and a KM analysis was used for several subgraphs to evaluate our model's ability to separate high-risk patients. Finally, as a proof of concept, we plotted one patient's node feature changes before and after 1 layer processing using a correlation heatmap, along with its neighbours' edge weights analysis to try to understand the inner workings of our model.

Results

Data Description

In the main cohort, we initially enrolled 2309 patients and after exclusion based on our criteria, a total of 1705 NSCLC patients were included in the study. The median age was 61 (interquartile range, 55-66 years). There were 1010 males (59·2%) and 695 women (40·8%). Tumours were more frequently located in the upper lobes (1018, 59·7%). A total of 1235 patients (72·4%) had adenocarcinoma, while 391 patients (22·9%) had squamous cell carcinoma. The distribution of pathologic stages was as follows: stage IA was present in 791 patients (46·4%), stage IB was present in 607 patients (35·6%), stage IIA was present in 133 patients (7·8%), and stage IIB was present in 174 patients (10·2%). The OS and RFS rates were 78·2 % (95% CI: 76·2% - 80·2%) and 74·2 % (70·8% -77·6%), respectively. The external validation dataset included a total of 127 patients of which 32 (25·2%) were females and 95 (74·8%) males, with a median age of 69 (interquartile range, 46-87 years). Upper lobe tumours were also more prevalent (76 patients, 59·8%). Among them were 95 patients diagnosed with adenocarcinoma and 30 with squamous cell carcinoma. The OS and RFS rates were 68·5 % (95% CI: 60·4 % - 77·7 %) and 59·1 % (95% CI: 50·4 % -67·8 %), respectively. Please refer to Table 1 for more detailed information.

211 Table 1: Feature distribution in the total patient cohorts, training and validation cohorts and the test cohorts

		TRAIN and VAL (n = 1492)	TEST (n = 213)		EXTERNAL (n= 127)	
Feature	Content	**Mean, SD, 95% CI / Count, %**		P	**Mean, SD, 95% CI / Count, %**	P
Age	Age	60·6, 8·7, (CI: 60·1, 61·0)	60·7, 9·5, (CI: 59·4, 62·0)	> 0·05	68·7, 9·1, (CI: 67·2, 70·1)	< 0·01**
Sex	Female No. (%); Male No. (%)	602 (33·3); 890 (66·7)	93 (33·3); 120 (66·7)	>0 ·05	32 (25·2); 95 (74·8)	< 0·01**
Resection	Sublobar Resection No. (%); Lobectomy No. (%); Bilobectomy No. (%); Pneumonectomy No. (%)	123 (8·2); 1292 (86·6); 59 (3·95); 18 (1·2)	23 (10·8); 180 (84·5); 7 (3·3); 3 (1·4)	> 0·05	/	/
Histology	Adenocarcinoma No. (%); Squamous Cell Carcinoma No. (%); Others No. (%)	1072 (71·4); 351 (23·5); 69 (4·6)	163 (76·5); 40 (18·8); 10 (4·7)	> 0·05	95 (74·8); 30 (23·6); 2 (1·6)	> 0·05
Tumour Location	LUL No. (%); LLL No. (%); RUL No. (%); RML No. (%); RLL No. (%)	384 (25·7); 211 (14·1); 504 (33·8); 146 (9·8); 247 (16·6)	51 (23·9); 37 (17·4); 79 (37·1); 15 (7·0) 31 (14·6)	>0 ·05	30 (23·6); 22 (17·3); 46 (36·2); 15 (11·8); 14 (11·0).	>0·05
Tumour Size	Tumour Size	2·68, 1·38, (CI: 2·61, 2·75)	2·55, 1·25, (CI: 2·38,2·71)	> 0·05	/	/
pTNM stage	Stage I No. (%); Stage II No. (%);	1219 (81·7); 273 (18·3)	179 (84·0); 34 (16·0)	> 0·05	97 (76·3); 30 (23·7)	< 0·01**
RFS Status	RFS No. (%)	1089 (73·0)	154 (72·3)	> 0·05	75 (59·1)	> 0·05
RFS Month	RFS Month	57·5, 24·5, (CI: 56·2, 58·7)	58·4, 23·4, (CI: 55·2, 61·5)	> 0·05	39·5, 26·9, (CI: 34·8, 44·2)	< 0·01**
OS Status	OS No. (survival %)	1166 (78·2)	167 (78·4)	> 0·05	87 (68·5)	>0·05
OS Month	OS Month	62·4, 19·9, (CI: 61·4, 63·4)	63·4, 18·4, (CI: 60·9, 65·9)	> 0·05	44·8, 27·8, (CI:40·9, 50·0)	< 0·01**

212

213

Model performance

To develop deep transformer graph learning–based biomarkers for overall survival prediction, we trained on the main cohorts, separated into training and validation datasets and then evaluated them separately on the testing set (213 patients) and the external set (127 patients). For OS prediction, our model achieved AUC values of 0·785 (95 % CI:0·716 - 0·855) and 0·695 (95 % CI:0·603 - 0·787) on the testing and external datasets, respectively, compared to 0·690 (95 % CI:0·600 - 0·780) and 0·634 (95 % CI: 0·544 - 0·724) for the TNM model, and 0·730 (95 % CI:0·640 - 0·820) and 0·626 (95 % CI:0·530 – 0·722) for ResNet-Graph model. For RFS prediction, our model achieved AUC values of 0·726 (95 % CI:0·653 - 0·800) and 0·700 (95 % CI:0·615 - 0·785) on the testing and external datasets, respectively, compared to 0·628 (95 % CI:0·542 - 0·713) and 0·650 (95 % CI: 0·561 - 0·732) for the TNM model, and 0·681 (95 % CI:0·598 - 0·764) and 0·595 (95 % CI:0·615 – 0·785) for ResNet-Graph model (Figure 3A and 3B). Additional survival analyses were performed using KM estimates for groups with low and high risk of mortality and recurrence, respectively, based on the median stratification of patient prediction scores (Figure 3C and 3D). All three models showed statistically significant differences in 5-year overall survival. For RFS prediction, the ResNet-Graph model was unable to distinguish between individuals at low and high risk ($p > 0·05$), while both Transformer-Graph and TNM models were able to separate high and low risk of recurrence-free survival groups ($p<0·05$). Additionally, the decision curve analysis (Figure 3E) and net benefit analysis (IDI, NRI) indicated that the Transformer-Graph model significantly outperformed the present TNM and ResNet-Graph models in terms of net benefit for both OS and RFS survival prediction.

As for detailed net benefit analysis, Transformer-Graph model outperformed the present TNM and ResNet-Graph models in terms of IDI and NRI. Our proposed model improved the survival prediction significantly compared with TNM regarding NRI (OS: 0·284, 95% CI: -0·112-0·519, $p<0·0001$; RFS:0·175, 95% CI:-0·115 - 0·486, $p<0·0001$) and IDI (OS:0·159, 95% CI: 0·103-0·214, $p=0·00032$; RFS:0·137, 95% CI: 0·086-0·189, $p=0·00074$). The results comparing with ResNet-Graph were reported in Supplementary IV.

Patients' clinical-based graph analysis

We visualised the whole internal set (Figure 4A) along with the testing cohorts subplot (Figure 4B) and analysed two challenging cases to better understand the population-based graph structure and how clinical data was integrated with node attributes (i.e. patients' tumour images). The testing subplot showed that while the graph structure (specified by the similarity score) was capable of broadly separating at-risk patients, several clusters had both high- and low-risk patients intermingled together, making them difficult to separate using traditional clinical information (see Figure 4C and 4D). The subsequent KM analysis indicated that by using Transformer-generated tumour attributes, high- and low-risk patients could be significantly discriminated.

Additionally, we analysed specifically as an example, patient No 44, and surrounding neighbours' edge weights distribution, as well as the initial and subsequent 1 layer node features. This patient was a high-risk patient who died after 38 months, with 42 neighbours. Initially, we analysed the correlation coefficient between neighbours' node features in order to determine the role that transformer-generated image features played prior to graph training. As

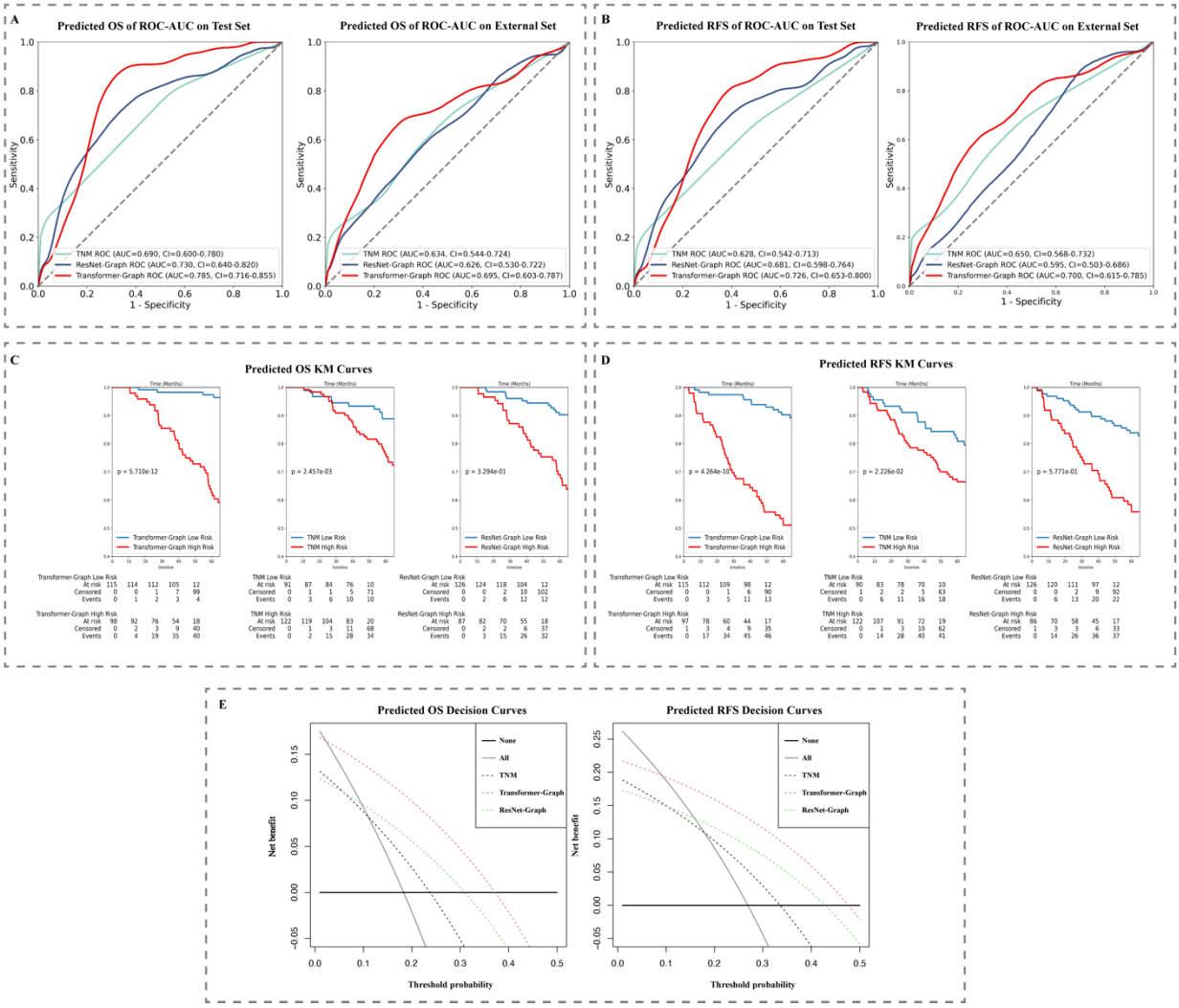


Figure 3. Model Performance: (A) ROC-AUC curve on test data and external set for OS and (B) RFS prediction and (C) KM curve on test data set for OS and (D) RFS prediction. (E) Decision curve on test data set for OS and RFS prediction.

illustrated in Figure 4E, the correlation matrix of Transformer-generated features revealed that almost all of patient No. 44's high-risk (dashed box nodes) and low-risk neighbours were highly correlated, implying that image features did not contain directly discriminative survival information before learning. We next then examined the distribution of neighbours' edge weights. As illustrated in Figure 4E, despite the fact that there were only five high-risk neighbours, the median value of similarity scores was slightly higher than that of low-risk neighbours (2·50 vs 2·00), indicating that the high-risk neighbour group was more closely connected to the target nodes from non-imaging information aspects.

After one layer of GraphSAGE updating, we discovered that the high-risk neighbours were more correlated with patient No. 44 (see Figure 4E GraphSAGE Layer 1, nodes in the dash boxes showed higher coefficient values), revealing that within our model, both neighbours' nodes and edge features contained survival-related information, and they contributed together to efficiently provide information for the target node learning.

Discussion

We demonstrated the feasibility of using Vision Transformer on CT images of the lung tumours to generate features for cancer survival analysis in this study. Additionally, we used a graph structure to embed patients' imaging and non-imaging clinical data separately in the graph neural network and attempted to explain how clinical data communicates with Transformer-generated imaging features for survival analysis. While Transformer and GNN models have been widely used in computer vision, their application in the medical field, particularly for survival prediction, is still evolving due to the complexity and unbalanced nature of medical data (high dimension, multiple data formats, including non-imaging data). In our study, we combined these two methods and created a specially designed graph structure to handle a variety of data formats, demonstrating the utility of Transformer-generated features in survival analysis and emphasizing the extent to which clinical data and imaging features contribute to the prediction. To our knowledge, this is the first work to demonstrate the feasibility of using Transformer in survival prediction using a graph data structure and exploratory analysis of the models' intuitions in an attempt to explain these state-of-the-art methods.

Our experiments indicated that the proposed model outperformed the commonly used TNM model in predicting survival not only on the testing dataset but also on the external dataset, despite the fact that the data distributions were significantly different (refer to Table 1, the survival distribution on the external dataset is significantly different from the internal dataset), demonstrating the model's generalisability for unseen data. The model's good performance indicated that both the Transformer-generated imaging features and the structure of our population graph (i.e., using graph edges and nodes to combine non-imaging clinical data and imaging data) contained useful information for survival. Additionally, the subplot graph on the testing dataset (Figure 4B) indicated that our graph structure was capable of approximate clustering high- and low-risk groups and segregating the majority of the high-risk patients. Meanwhile, when patients were similar in terms of demographic information and it was hard to determine the risk patients by traditional clinical methods (refer to Figures 4C and 4D the dense graphs containing both pink and blue nodes), the Transformer-generated image features and edge weights had more roles to play in determining the

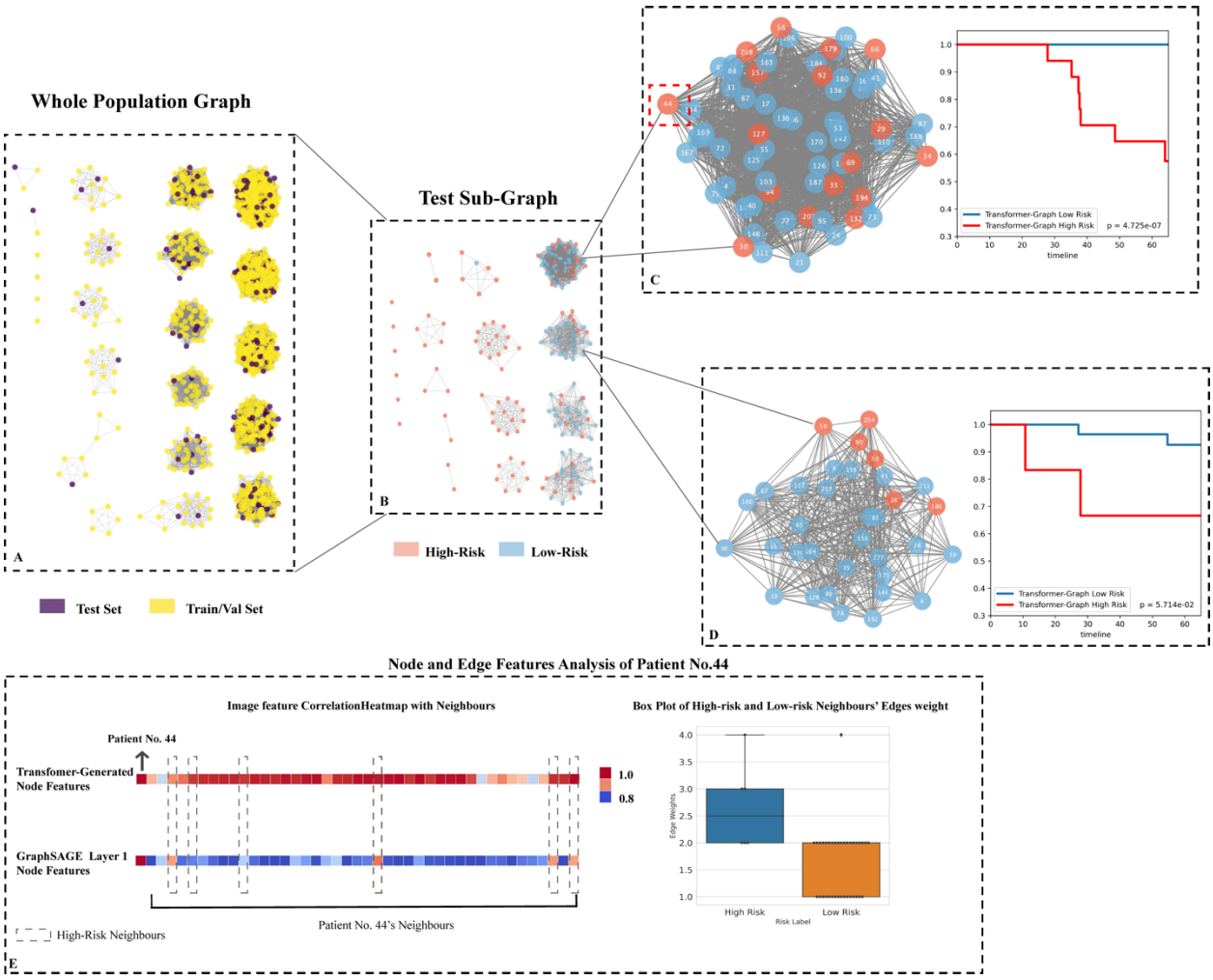


Figure 4. Testing set graph analysis. (A) A visual representation of the whole cohort population graph of 1705 patients. (B) A visual representation of the testing sub-graph of 213 patients. (C) and (D) Two subgraphs containing challenging cases where the graphs contained both high- and low-risk patients. (E) Node features' correlation heatmaps and edge weights distribution of patient No. 44: Each square represents a neighbour's node features' correlation coefficient, higher values (red colour) reveal closer relation with the target node; The box plot of 42 neighbours indicates that the high-risk neighbours (blue box) have higher edge weights median.

differences between neighbours. More specifically, the Transformer-generated features did not contain directly discriminative survival information before learning, while with edge weights together, effective information from with neighbours' node features could be passed. In this case, all patient's node features could be effectively updated, and high-risk patients could be better discriminated as in Figure 4E.

Our study contains several strengths. First, our dataset is relatively large, encompassing both contrast and non-contrast CT. This not only aided in the model's generalisation learning but also allows for flexibility in the imaging standards in clinical settings. Second, our graph model demonstrated the ability to combine non-imaging clinical features with imaging features in an understandable manner, implying a new direction of embedding multi-data with deep learning models. Finally, we sought to understand the roles of imaging and non-imaging features in determining high-risk nodes within the graph neural network, which could aid clinicians in comprehending the internal workings of the neural networks.

There are some limitations worth noting. First, whilst the proposed model significantly outperformed the TNM model on the external dataset (OS prediction AUC 0·693 vs 0·633, RFS prediction RFS 0·700 vs 0·650), the model's performance on the external set was below that of the testing set (AUC 0·783 and 0·726 for OS and RFS). One reason could be that the patients' demographics were different, particularly in terms of age (the external group's average age was ten years older than the main cohort), cancer staging (84·0 % stage I in the main cohort while 76·3 % in the external testing set), and gender (male percentage 66·7 % vs 78·3 %). Given the fact that the two datasets originate from distinct countries, as well as the differences in ethnicity, treatment and follow-up strategies (see Table 1, especially the mean follow-up time) may also have an impact on the prediction performance. Second, the initial step requires the human observer to identify the tumour and draw a bounding box which in our study was still a manual procedure. As the pipeline for automatic tumour detection and segmentation becomes more mature, this step can potentially be automated allowing for ease of translation into the clinics.

In conclusion, the population graph deep learning model constructed using Transformer-generated imaging and non-imaging clinical features was proven to be effective at predicting survival in patients with early-stage lung cancer. The subanalysis concluded that by developing a meaningful similarity score function and comparing patients' non-imaging characteristics such as age, gender, histology type, and tumour location, the majority of high-risk patients can already be separated. Additionally, when high- and low-risk patients shared very similar demographic information, TNM information provided additional information for survival prediction when combined with tumour imaging features.

Conflict of interest statements

All authors declare no competing interests.

Data Availability

The current manuscript is a computational study, so no data have been generated for this manuscript.

Code Availability

To aid reproducibility of research, our codes are published on the Github repository:
https://github.com/SereneLian/TransGNN-Lung

Reference

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin*. Jan 2021;71(1):7-33. doi:10.3322/caac.21654

2. Ambrogi MC, Fanucchi O, Cioni R, et al. Long-term results of radiofrequency ablation treatment of stage I non-small cell lung cancer: a prospective intention-to-treat study. *Journal of Thoracic Oncology*. 2011;6(12):2044-2051.

3. Goldstraw P, Chansky K, Crowley J, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. *Journal of Thoracic Oncology*. 2016;11(1):39-51.

4. Giroux DJ, Van Schil P, Asamura H, et al. The IASLC lung cancer staging project: a renewed call to participation. *Journal of Thoracic Oncology*. 2018;13(6):801-809.

5. Malone ER, Oliva M, Sabatini PJ, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. *Genome medicine*. 2020;12(1):1-19.

6. Du R, Lee VH, Yuan H, et al. Radiomics model to predict early progression of nonmetastatic nasopharyngeal carcinoma after intensity modulation radiation therapy: a multicenter study. *Radiology: Artificial Intelligence*. 2019;1(4):e180075 2638-6100.

7. Aonpong P, Iwamoto Y, Wang W, Lin L, Chen Y-W. Hand-Crafted and Deep Learning-Based Radiomics Models for Recurrence Prediction of Non-Small Cells Lung Cancers. *Innovation in Medicine and Healthcare*. Springer; 2020:135-144.

8. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature Reviews Clinical Oncology*. 2021:1-15.

9. Carmody DP, Nodine CF, Kundel HL. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*. 1980;9(3):339-344.

10. Chirra P, Leo P, Yim M, et al. Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI. *Journal of Medical Imaging*. 2019;6(2):024502.

11. Mirsadraee S, Oswal D, Alizadeh Y, Caulo A, van Beek E, Jr. The 7th lung cancer TNM classification and staging system: Review of the changes and implications. *World J Radiol*. 2012;4(4):128-134. doi:10.4329/wjr.v4.i4.128

12. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339

13. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*. 2018;392(10162):2388-2396. doi:10.1016/S0140-6736(18)31645-3

14. Nabulsi Z, Sellergren A, Jamshy S, et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. *Scientific Reports*. 2021/09/01 2021;11(1):15523. doi:10.1038/s41598-021-93967-2

15. Xu Y, Hosny A, Zeleznik R, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin Cancer Res*. 2019;25(11):3266-3275. doi:10.1158/1078-0432.CCR-18-2495

16. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*. 2021;

17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017:5998-6008.

18. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:201011929*. 2020;

19. Zhou L, Liu H, Bae J, He J, Samaras D, Prasanna P. Self Pre-training with Masked Autoencoders for Medical Image Analysis. *arXiv preprint arXiv:220305573*. 2022;

20. Holzinger A, Haibe-Kains B, Jurisica I. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*. 2019;46(13):2722-2730.

21. Xue Y, Xu T, Long LR, et al. Multimodal recurrent model with attention for automated radiology report generation. Springer; 2018:457-466.

22. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*. 2019;49(1):15-21.

23. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:160902907*. 2016;

24. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017; Long Beach, California, USA.

25. Bakr S, Gevaert O, Echegaray S, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*. 2018;5(1):1-9.

26. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic resonance imaging*. 2012;30(9):1323-1341.

27. Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:210410972*. 2021;

28. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*. 2018;18(1):1-12.

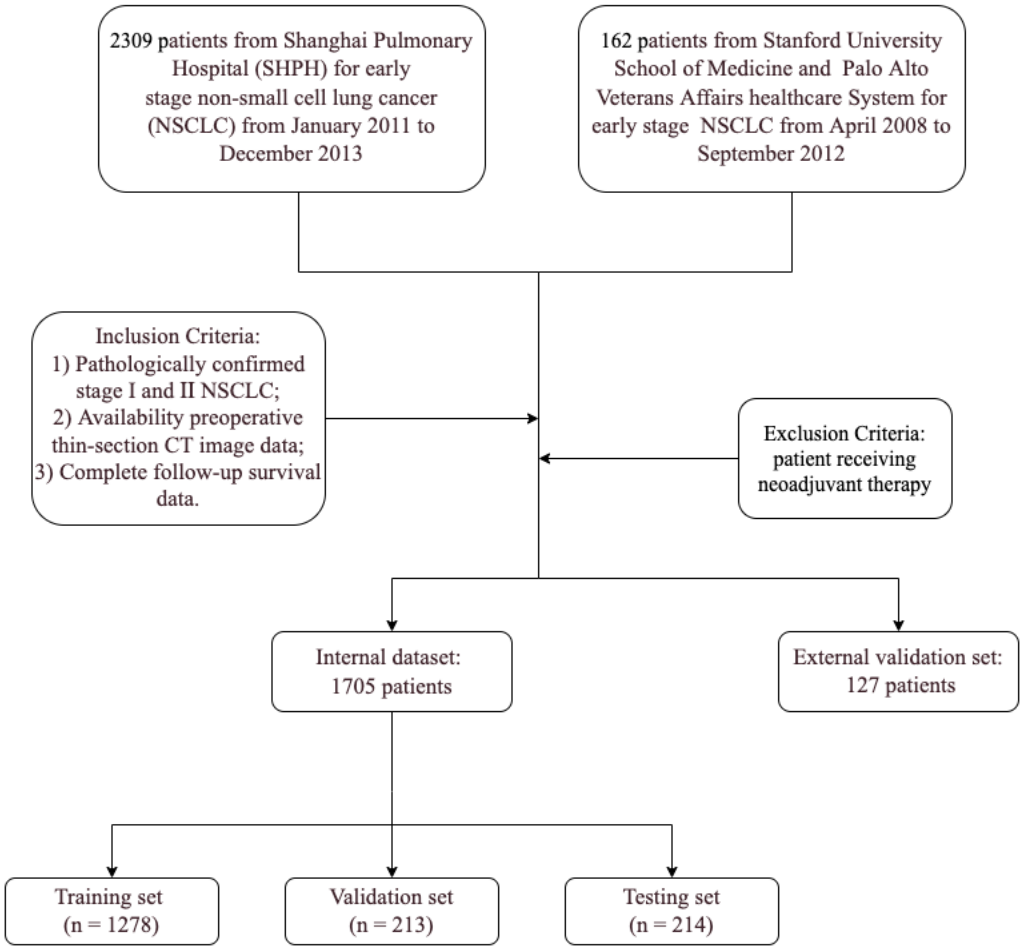
412 29. Khanna A, Londhe ND, Gupta S, Semwal A. A deep Residual U-Net convolutional neural network for
413 automated lung segmentation in computed tomography images. *Biocybernetics and Biomedical Engineering*.
414 2020/07 2020;40(3):1314-1327. doi:10.1016/j.bbe.2020.07.007

415 30. Chen S, Ma K, Zheng Y. Med3D: Transfer Learning for 3D Medical Image Analysis.
416 2019:arXiv:1904.00625. Accessed April 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190400625C

417

418

Supplementary I



Supplementary Figure 1: Overall flow of the study in both internal and external dataset

Supplementary II Scanner parameter and follow-up strategies

CT scans ranged from thoracic inlet to subcostal plane and were obtained before surgical resection from 2 CT machines: Brilliance (Philips Medical Systems Inc, Cleveland, OH) and SOMATOM Definition AS (Siemens Aktiengesell-schaft, Munich, Germany).

CT parameters of Brilliance (Philips Medical Systems Inc) were as follows: 64 x 1 mm acquisition; 0.75-second rotation time; slice width 1 mm; tube voltage, 120 kVp; tube current, 150 to 200 mA; lung window center: -700 Hounsfield units (HU), and window width:1200 HU; mediastinal window center: 60 HU and window width: 450 HU level; pitch: 0.906; and field of view (FOV): 350 mm.

CT parameters of the SOMATOM Definition AS (Siemens Aktiengesell-schaft) were as follows: 128 x 1 mm acquisition; 0.5-second rotation time; slice width: 1 mm; tube voltage: 120 kVp; tube current: 150 to 200 mA; lung window center: -700 HU and window width 1200 HU; and mediastinal window center: 60 HU and window width: 450 HU level; FOV: 300 mm; pitch: 1.2; and FOV: 350 mm. CT images were reconstructed into 0.67- to 1.25-mm section thicknesses according to a high-resolution algorithm.

Follow-up was conducted through outpatient examinations or telephone calls.

Chest CT scan and abdominal ultrasound/CT were performed on follow-up visits within a duration of 3, 6, and 12 months after operation and annually thereafter for 5 years. Magnetic resonance imaging for brain and bone scan were annually performed for 5 years or when the patient had signs or symptoms of recurrence.

Supplementary III Similarity Score Definition

Similarity score for patient x and patient y:

$$Sim(x, y) = C_{xy} * L_{xy} * H_{xy} * T_{xy}$$

C_{xy}: if x and y have same gender, get 1 point; if x and y's age difference is within 5 year, get another 1 point.

L_{xy}: if x and y's tumours locate at the same lung lobes, get 1 point.

H_{xy}: if x and y's histology of tumours is the same type, get 1 point.

T_{xy}: if x and y have the same T stage, get 1 point; if x and y have the same N stage, get another point; if x and y have the same M stage, get another 1 point.

When $Sim(x, y)$>0, patient x and y can be connected.

Supplementary IV: ResNet-Graph NRI and IDI results

Transformer-Graph comparing with ResNet-Graph, regarding NRI (OS:0.240, 95% CI: -0.325-0.600, P< .001; RFS: 0.104, 95% CI: -0.41-0.389, P< .001) and IDI (OS:0.075 , 95% CI: 0.068 – 0.082 , P< .05; RFS: 0.063, 95% CI: 0.027 -0.098, P< .05).