SCIENTIFIC REPORTS

A Large-scale Cross-sectional Study of ALK Rearrangements and EGFR Mutations in Non-small-cell Lung Cancer in Chinese Han Population

Shaodong Hong[1,2,3]*, Wenfeng Fang[1,2,3]*, Zhihuang Hu[1,2,3]*, Ting Zhou[1,2,3], Yue Yan[1,2,3], Tao Qin[1,2,3], Yanna Tang[4], Yuxiang Ma[1,2,3], Yuanyuan Zhao[1,2], Cong Xue[1,2,3], Yan Huang[1,2,3], Hongyun Zhao[1,2,3] & Li Zhang[1,2,3]

[1]State Key Laboratory of Oncology in South China, Guangzhou, China, [2]Collaborative Innovation Center for Cancer Medicine, Guangzhou, China, [3]Department of Medical Oncology, Sun Yat-sen University Cancer Center, Guangzhou, China, [4]Department of Oncology, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, China.

The predictive power of age at diagnosis and smoking history for *ALK* rearrangements and *EGFR* mutations in non-small-cell lung cancer (NSCLC) remains not fully understood. In this cross-sectional study, 1160 NSCLC patients were prospectively enrolled and genotyped for *EML4-ALK* rearrangements and *EGFR* mutations. Multivariate logistic regression analysis was performed to explore the association between clinicopathological features and these two genetic aberrations. Receiver operating characteristic (ROC) curves methodology was applied to evaluate the predictive value. We showed that younger age at diagnosis was the only independent variable associated with *EML4-ALK* rearrangements (odds ratio (OR) per 5 years' increment, 0.68; $p < 0.001$), while lower tobacco exposure (OR per 5 pack-years' increment, 0.88; $p < 0.001$), adenocarcinoma (OR, 6.61; $p < 0.001$), and moderate to high differentiation (OR, 2.05; $p < 0.001$) were independently associated with *EGFR* mutations. Age at diagnosis was a very strong predictor of *ALK* rearrangements but poorly predicted *EGFR* mutations, while smoking pack-years may predict the presence of *EGFR* mutations and *ALK* rearrangements but with rather limited power. These findings should assist clinicians in assessing the likelihood of *EML4-ALK* rearrangements and *EGFR* mutations and understanding their biological implications in NSCLC.

Non-small-cell lung cancer (NSCLC) is a leading cause of cancer mortality[1]. Although chemotherapy remains the mainstream treatment of advanced NSCLC, small molecular tyrosine kinase inhibitors (TKIs) targeting specific driver mutations have resulted in favorable response rate, progression-free survival (PFS), and quality of life in sensitive population[2–7].

Epidermal growth factor receptor (EGFR) mutations are the first discovered druggable targets in NSCLC[8]. Two classes of *EGFR* mutations, exon 19 deletions and exon 21 substitutions, account for the majority of *EGFR* mutations reported (~90%)[9]. These mutations are correlated with better response to gefitinib, erlotinib and afatinib[2–4]. And they were more frequently observed in Asian population, never smokers, females, and patients with adenocarcinoma[10]. A reverse relationship between cumulative smoking pack-years and the frequency of *EGFR* mutations has widely been reported[11,12], suggesting some predictive value of smoking status for the presence of *EGFR* mutations. However, the association between age at diagnosis and *EGFR* mutations remains controversial and little data is available regarding the predictive value of age at diagnosis on *EGFR* mutations[10,13–15].

Fusion of the *Echinoderm microtubule-associated protein like-4* (*EML4*) and *anaplastic lymphoma kinase* (*ALK*) represents another distinct mechanism of driver mutation in NSCLC. The fusion protein is highly oncogenic both in vitro and vivo, resulting in constitutive ALK pathway activation and ultimately cancer development[16,17]. Several clinical trials have demonstrated the remarkable efficacy of crizotinib for the treatment of metastatic NSCLC patients who harbor ALK rearrangements, which led to its approval from the US Food and Drug Administration and the European Medicines Agency[6,18,19]. *EML4-ALK* rearrangements were mostly reported to be associated with younger age at diagnosis and adenocarcinoma[20–22]. However, inconsistent reports

still exist[6,22,23]. Also, the accuracy of age at diagnosis and smoking status in predicting *EML4-ALK* rearrangements has not been established.

One of the most challenging problems in clinical practice is to acquire adequate tumor tissue for genomic analysis. Therefore, using available clinicopathological data to predict the likelihood of certain genetic aberrations is of special significance. Furthermore, *EML4-ALK* rearrangements and *EGFR* mutations represent two distinct oncogenic mechanisms, which might have different clinicopathological features. However, a limited number of studies have concerned about such difference in a single dataset.

We therefore carried out this epidemiological study based on large-scale genotyped NSCLC patients to evaluate the distinct clinicopathological features associated with *ALK* rearrangements and *EGFR* mutations in Chinese Han population, as well as the predictive value of age at diagnosis and smoking pack-years on these two genetic aberrations.

Results

Population characteristics. From 10th January 2012 to 25th April 2014, 1377 NSCLC patients were prospectively enrolled in Sun Yat-sen University Cancer Center. After excluding 102 patients who had insufficient tumor tissue for genomic analysis and another 115 patients who refused to participate in the study, a total of 1160 patients were finally included. Figure 1 outlines the process of patient selection. Overall clinicopathological features of the included patients are summarized in Table 1. The median age at diagnosis was 57 years (range: 19–85 years). Among them, 39.1% were females, 54.0% were never-smokers, 78.1% had adenocarcinoma, and 43.0% were diagnosed at stage IIIB-IV. Females were more likely to be non-smokers ($p < 0.001$). The incidence of *EML4-ALK* rearrangements and *EGFR* mutations was 8.1% ($n = 94$) and 33.8% ($n = 392$), respectively. Two cases of concurrent ALK rearrangement and EGFR mutation were noted, whose clinical and pathological data are presented in Supplementary Table S1 online.

Association between clinicopathological data and *EML4-ALK* rearrangements. Patients with *EML4-ALK* rearrangements were significantly younger at diagnosis than those without such rearrangements (median age, 45 versus 58 years; $p < 0.001$). Never smokers were more likely to harbor *EML4-ALK* rearrangements than smokers (10.1% versus 6.1%; $p = 0.005$). Patients with advanced NSCLC (stage IIIB-IV) had significantly higher incidence of *EML4-ALK* rearrangements compared with those diagnosed at stage I-IIIA (10.6% versus 6.2%; $p = 0.006$). 8.9% of adenocarcinoma and 4.7% of non-adenocarcinoma had *EML4-ALK* rearrangements, respectively ($p = 0.036$). For symptoms at first onset, none of the symptoms were found to be significantly associated with *EGFR* mutations or *ALK* rearrangements. The results of univariate logistic analysis are shown in Table 1.

In order to adjust for confounding factors, we carried out multivariate logistic regression analysis. The results showed that only younger age at diagnosis remained independently associated with *EML4-ALK* rearrangements (odds ratio (OR) per 5 years' increment, 0.68; 95% confidence interval (CI), 0.62–0.75; $p < 0.001$). The remaining variables, including smoking status ($p = 0.223$), cancer stage ($p = 0.500$), and pathological types ($p = 0.051$) were no longer associated with *EML4-ALK* rearrangements.

Association between clinicopathological data and *EGFR* mutations. The results of univariate analysis are shown in Table 1. Female patients were more likely to have *EGFR* mutations compared with male patients (46.7% versus 25.5%; $p < 0.001$). Never smokers had higher incidence of *EGFR* mutations than smokers did (45.9% versus 21.1%; $p < 0.001$). Adenocarcinoma ($p < 0.001$) and moderate to high differentiation ($p < 0.001$) were also significantly associated with *EGFR* mutations. Subsequent multivariate logistic regression analysis showed that lower tobacco exposure (OR per 5 pack-years' increment, 0.88; 95% CI, 0.85–0.92; $p < 0.001$), adenocarcinoma (OR, 6.61; 95% CI, 3.58–12.19; $p < 0.001$), and moderate to high differentiation (OR, 2.05; 95% CI, 1.55–2.71; $p < 0.001$) were independent predictors of harboring *EGFR* mutations. However, gender ($p = 0.154$) and cancer stage ($p = 0.767$) were not considered to be independently associated anymore.

Age at diagnosis as a predictor of *EML4-ALK* rearrangements and *EGFR* mutations. We sought to investigate the impacts of age at diagnosis (denoted as "age" in the following contents) on the incidence of *EML4-ALK* rearrangements in detail, compared with those on *EGFR* mutations. The incidence of *EML4-ALK* rearrangements decreased dramatically with increasing age, while the incidence of *EGFR* mutations increased with age till 50–59 years and remained nearly unchanged with age (Figure 2). Patients under the age of 30 had a 44% (7/16) incidence of *EML4-ALK* rearrangements, compared with a 4% (5/135) incidence in those above 70 years ($p < 0.001$). Notably, when stratified by gender or smoking status, similar age-distribution patterns were also observed (Figure 3).

To seek for cut-off value of age as a predictor of *EML4-ALK* rearrangements, ROC curve was plotted. The area under ROC curve (AUC) was 0.74 (95% CI, 0.68–0.80) with the cut-off age of 50.5 years (sensitivity, 73%; specificity, 70%) (Figure 4A). Patients under 50.5 years old had an 18.5% (66/356) incidence of *EML4-ALK* rearrangements compared with a 3.5% (28/804) incidence in patients above 50.5 years old (OR = 6.1; $p < 0.001$). This cut-off value also showed fair discriminative power in patients with different clinicopathological features (Table 2). In patients younger than 50.5 years old, when we added "*EGFR* wild type" and "adenocarcinoma" into the enrichment strategy, we got a 29.4% (58/197) incidence of *EML4-ALK* rearrangements.

We also plotted ROC curve for age as a predictor of *EGFR* mutations (Figure 4B). The AUC was only 0.52 (95% CI, 0.49–0.56).

Smoking pack-years before diagnosis as a predictor of *EML4-ALK* rearrangements and *EGFR* mutations. The incidence of *EML4-ALK* rearrangements and *EGFR* mutations by smoking pack-years was shown in Figure 5. Briefly, the incidence of *EGFR* mutations decreased with increasing smoking pack-years. A 5 pack-years' increment led to a 12% decrease in the likelihood of *EGFR* mutations. However, there was a plateau of the incidence of *EGFR* mutations after more than 10 pack-years of cigarettes were consumed. Even in patients with more than 80 smoking pack-years, the incidence of *EGFR* mutations was as high as 13.2% (5/38). As for *EML4-ALK* rearrangements, the incidence peaked at
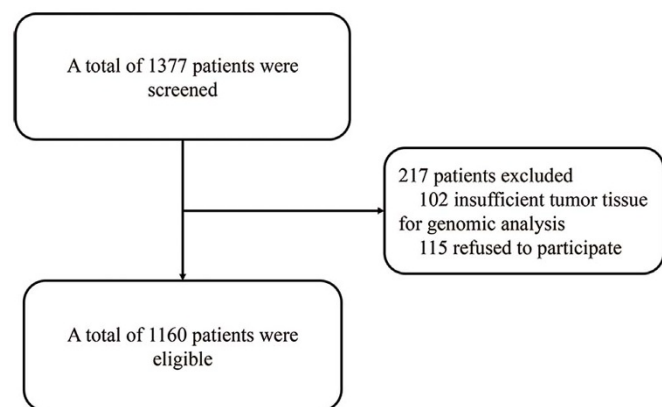

Figure 1 | Flow diagram of patient selection process.

Table 1 | Demographic and clinicopathological characteristics of non-small-cell lung cancer patients with defined *EML4-ALK* rearrangement and *EGFR* mutation status

Characteristics	Total, N (%)	*EML4-ALK* rearrangement			*EGFR* mutation		
		Positive	Negative	p	Positive	Negative	p
Patients, *n*	1160	94	1066		392	768	
Age at diagnosis, years							
Median	57	45	58	<**0.001**	57	57	0.597
Range	19–85	20–81	19–85		27–81	19–85	
Gender, *n*							
Female	454 (39.1)	40	414	0.479	212	242	<**0.001**
Male	706 (60.9)	54	652		180	526	
Smoking status, *n*							
Never smoking	626 (54.0)	64	562	**0.005***	272	320	<**0.001***
Smoking	525 (45.3)	30	495		118	441	
Unknown	9 (0.8)	0	9		2	7	
Pathology, *n*							
Adenocarcinoma	906 (78.1)	81	825	**0.036**§	368	538	<**0.001**§
Squamous	168 (14.5)	9	159		12	156	
Adenosquamous	22 (1.9)	2	20		9	13	
Others	64 (5.5)	2	62		3	61	
Differentiation, *n*							
Moderate to high	434 (37.4)	31	403	0.649£	200	234	<**0.001**£
Poor	632 (54.5)	55	577		161	471	
Undetermined	94 (8.1)	8	86		31	33	
Stage, *n*							
I–IIIA	661 (57.0)	41	620	**0.007**	212	449	0.154
IIIB–IV	499 (43.0)	53	446		180	319	
Family history of cancer, *n*							
Yes	238 (20.5)	22	216	0.470	79	159	0.826
No	922 (79.5)	72	850		313	609	
Cough, n							
Yes	856 (73.8)	68	788	0.738	300	488	0.139
No	304 (26.2)	26	278		92	186	
Chest pain, n							
Yes	504 (43.4)	39	465	0.690	181	323	0.181
No	656 (56.6)	55	601		211	445	
Hemoptysis, n							
Yes	258 (22.2)	14	244	0.077	93	165	0.867
No	902 (77.8)	80	822		299	603	
Shortness of breath, n							
Yes	336	20	316	0.089	121	215	0.308
No	824	74	750		271	553	

*p values compare never smoking with smoking.
§p values compare adenocarcinoma plus adenosquamous with squamous and others.
£p values compare moderate to high differentiation with poor differentiation.
Bold fonts indicate significant p values.

0–10 pack-years (20%) and then dropped with increasing cigarettes smoking.

The AUC for smoking pack-years as a predictor of *EML4-ALK* was 0.60 (95% CI, 0.55–0.65) with the optimal cut-off value of 10.25 pack-years (sensitivity, 41%; specificity, 82%) (Figure 6A). Patients who smoked less than 10.25 pack-years were more likely to harbored *EML4-ALK* compared to those who smoked more than 10.25 pack-years (11.1% versus 3.8%; $p < 0.001$). For *EGFR* mutations, the AUC was 0.66 (95% CI, 0.63–0.70) with the cut-off value of 2.75 pack-years (sensitivity, 55%; specificity 77%) (Figure 6B). Patients who smoked less than 2.75 smoking pack-years had a 45.9% (291/634) incidence of *EGFR* mutations compared with a 17.4% (87/500) incidence in those who smoked more than 2.75 pack-years (OR, 4.0; 95% CI, 3.0–5.3; $p < 0.001$).

Discussion

In clinical practice, the discrimination of *ALK* rearrangements and *EGFR* mutations in NSCLC has critical therapeutic implications. *EGFR* mutations confer sensitivity to EGFR TKIs while patients with *ALK* rearrangements response well to ALK TKIs. However, *ALK* rearrangements are associated with resistance to EGFR TKIs[21,24]. Due to some shared features of *EML4-ALK* rearrangements and *EGFR* mutations such as adenocarcinoma histology and never/light smokers, it is important to investigate other distinct features of these two genetic aberrations. To our knowledge, this is the first study to investigate the roles of patients' clinicopathological features in predicting the presence of *EML4-ALK* rearrangements and *EGFR* mutations.

We found age at diagnosis was the only variable that independently predicted *EML4-ALK* rearrangements. There was an inverse relationship between age at diagnosis and the frequency of *EML4-ALK* rearrangements. A 5-year's increment in age decreased the likelihood of *EML4-ALK* rearrangements by 32%. Using ROC curve, the cut-off age at diagnosis for predicting *EML4-ALK* rearrangements was determined to be 50.5 with a sensitivity of 73% and a specificity of 70%. Patients younger than 50.5 years had a five-fold greater chance of harboring *EML4-ALK* rearrangements compared with those older than 50.5 years. Noteworthy, this cut-off value also shows fair discriminative power in patients with various clinicopathological features (Table 2). For male and female patients, the
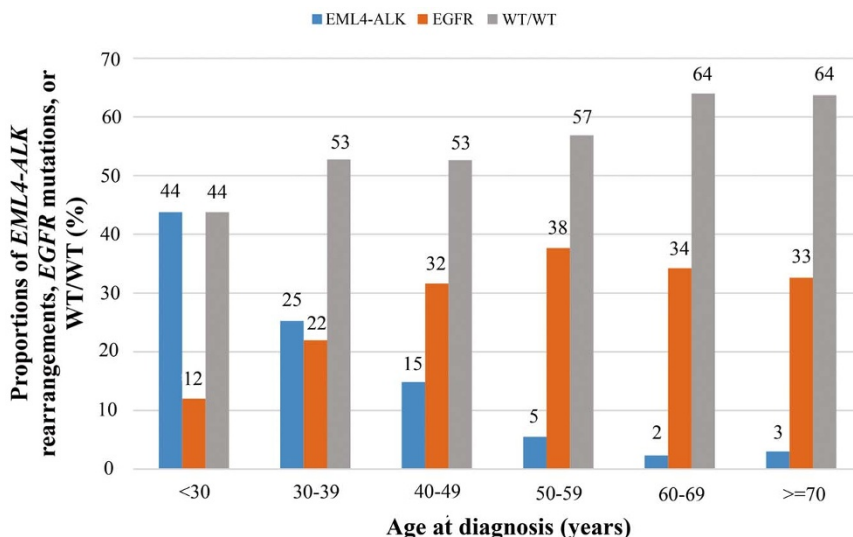
Figure 2 | The incidence of *EML4-ALK* rearrangements, *EGFR* mutations, and *WT/WT* in non-small-cell lung cancer patients according to different age groups (at diagnosis). *WT/WT*, wild type *ALK* and *EGFR*. There is an inverse relationship between age at diagnosis and the incidence of *EML4-ALK* rearrangements.
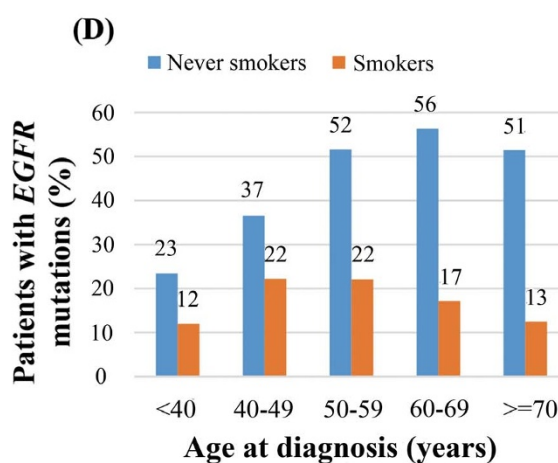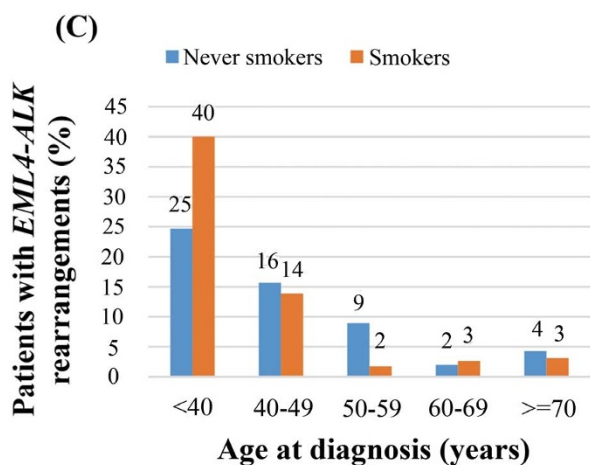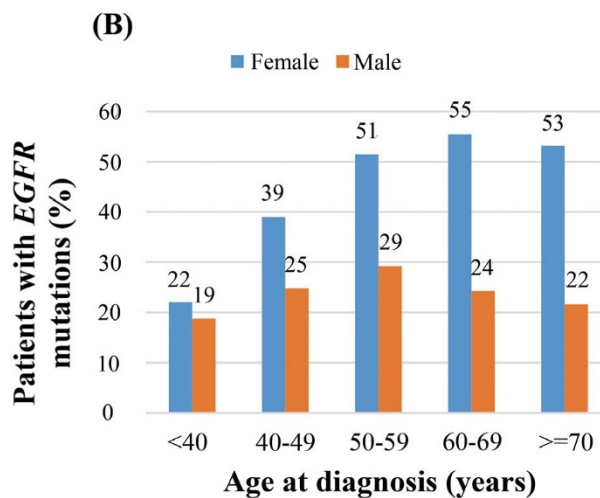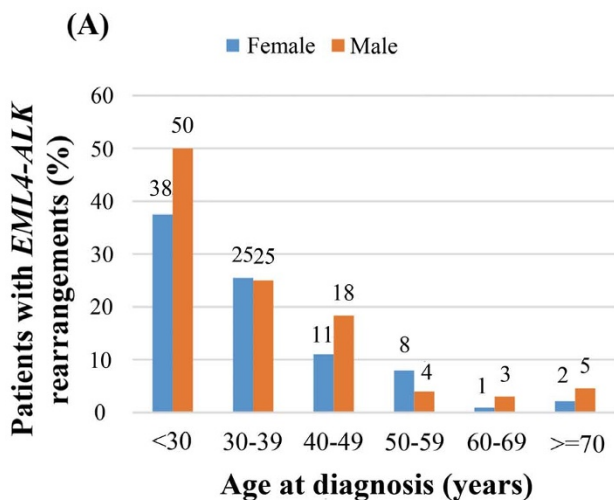

Figure 3 | Age distribution (at diagnosis) of *EML4-ALK* rearrangements and *EGFR* mutations in non-small-cell lung cancer patients at diagnosis stratified by (A) & (B) gender and (C) & (D) smoking status.
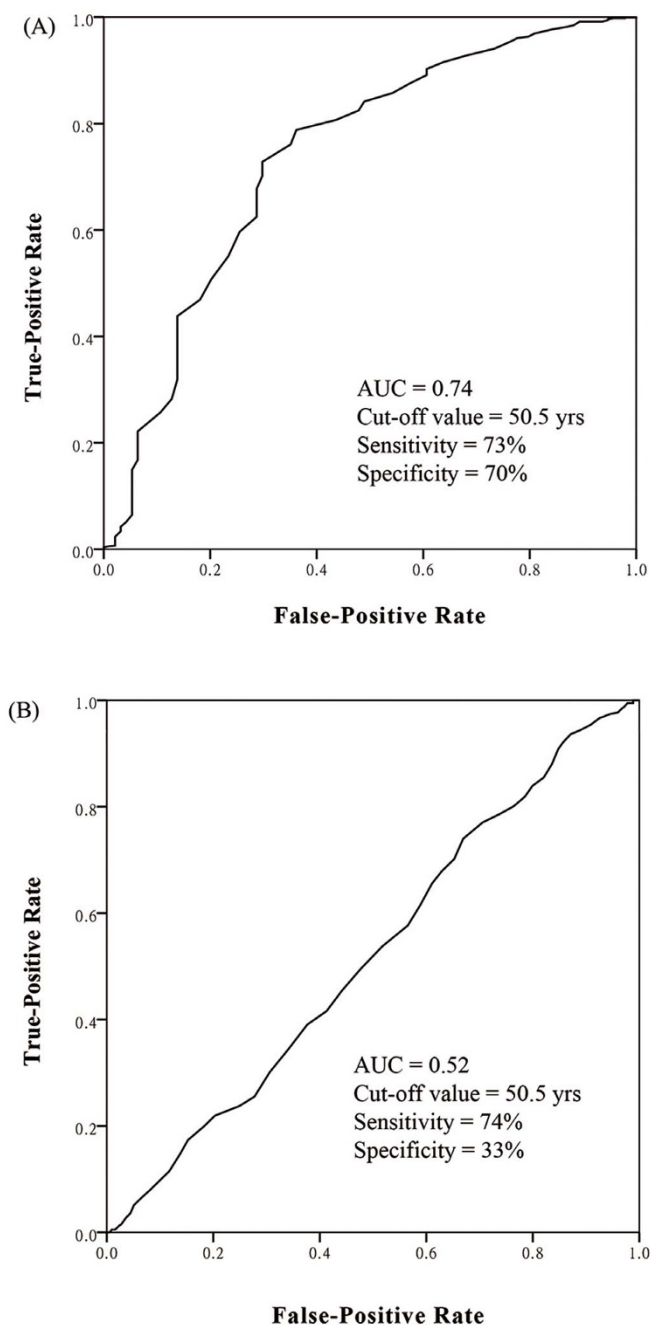
Figure 4 | Receiver operating characteristics (ROC) curves for age at diagnosis as predictors of (A) *EML4-ALK* rearrangements and (B) *EGFR* mutations in non-small-cell lung cancer. The optimal cut-off value is the point closest the upper-left corner of the graph. AUC, areas under ROC curve.

ORs of ALK rearrangement in patients younger than 50.5 years versus those older than 50.5 years were similar (OR = 6.1 and 6.8, respectively). This implies that gender difference in ALK rearrangement might be limited. However, in non-smokers, the discriminative power seems lower (OR = 2.6). One possible explanation is that smoking status, though not as remarkable as age at diagnosis, still affects the incidence of ALK rearrangement. For histological subtypes (adenocarcinoma versus non-adenocarcinoma) and cancer stage (I-IIIA versus IIIB-IV), age at diagnosis also satisfactorily predict the likelihood of ALK rearrangement. Taken together, these results indicate age at diagnosis alone is a very strong predictor of *ALK* rearrangements in NSCLC. However, attention should be paid

to non-smokers older than 50.5 years old who still have 7.6% incidence of ALK rearrangement. In an enriched population (younger than 50.5 years old, *EGFR* wild type and with adenocarcinoma histology), we found one-third of patients harbored *EML4-ALK* rearrangements. This enrichment strategy is useful when assessing the likelihood of *EML4-ALK* rearrangements in NSCLC patients. Our results have several conflicts with a prospective *ALK* screening study, which showed that male patients, light/never smokers and N3 stage were independently associated with *ALK* rearrangements[22]. The discrepancy, which remains to be elucidated, is probably due to the small sample size in the previous study (only 16 positive cases from 116 patients), the obvious selection bias reported, or simply the ethnic difference.

Our results also revealed that the predictive power of smoking pack-years on *EML4-ALK* rearrangements was limited. This is unsurprising since smoking status was not an independent variable associated with *EML4-ALK* rearrangements. Indeed, they have been sporadically reported in both smokers and nonsmokers[22,25,26], indicating a lack of association between smoking status and *EML4-ALK* rearrangements.

As for *EGFR* mutations, we found that lower tobacco exposure, adenocarcinoma, and moderate to high differentiation were independently associated with *EGFR* mutations. Similar to a prospective *EGFR* screening study in Asian[10], we found female was associated with *EGFR* mutations in univariate analysis but not in multivariate logistic regression model, suggesting internal association between gender and other variables such as smoking status. Until now, it is widely believed that female rather than male patients should be tested for *EGFR* mutations. This unspoken prejudice may miss out a substantial of patients who will benefit from targeted therapy. A previous study which established a nomogram to predict the presence of *EGFR* mutations also indicated that gender has little contribution to such prediction while smoking pack-years is the strongest predictor[27].

The association between age at diagnosis and *EGFR* mutations has long been controversial. Some studies showed that patients with *EGFR* mutations were older than those without *EGFR* mutations[13–15], While other studies found no significant association[10,28]. In the present study, we found age at diagnosis was not associated with *EGFR* mutations, which was further supported by ROC curve (AUC = 0.52). These results suggest that the likelihood of *EGFR* mutations is poorly predicted by age at diagnosis. One possible explanation might be that there is a peak incidence of *EGFR* mutations around 60 years old (Figure 1, "n" shape distribution). This was also supported by the nomogram model from Girard et al's study which shows that patients aged between 60 and 70 have higher probability of *EGFR* mutations[27].

Finally, we explored the roles of smoking pack-years in predicting *EGFR* mutations. We found the incidence of *EGFR* mutations was inversely related to smoking pack-years, similar to a previous study[11]. In that study, they concluded that smoking pack-years strongly predicted *EGFR* mutations (AUC = 0.78). No patients that have smoked more than 75 pack-years harbor *EGFR* mutations. In the current study, however, the AUC was lower (AUC = 0.66). Indeed, we observed a plateau of the incidence of *EGFR* mutations after more than 10 pack-years of cigarettes consumption, suggesting that smoking has no cumulative effects on *EGFR* mutations. Patients who have smoked more than 80 pack-years still had a 13.2% incidence of *EGFR* mutations. Therefore heavier smokers should still be considered for *EGFR* mutation tests[29]. This finding also explains why the predictive accuracy of smoking pack-years reported here was lower than previously believed[11].

The distinct age-distribution patterns of *EML4-ALK* rearrangements and *EGFR* mutations may also imply the difference of oncogenic potency. NSCLCs with *EGFR* mutations are generally dormant and would take a longer time to become clinically detectable, resulting in accumulated occurrence of *EGFR* mutant tumors in patients of

Table 2 | Odds ratio of *EML4-ALK* rearrangements in patients younger than 50.5 yrs versus patients older than 50.5 yrs, stratified by clinicopathological features

Characteristics	<50.5 years old		>50.5 years old		OR	95% CI	p
	Patients with *EML4-ALK* rearrangements/total	Incidence, %	Patients with *EML4-ALK* rearrangements/total	Incidence, %			
Gender							
Male	35/186	18.8	19/520	3.7	6.1	3.4–11.0	<0.001
Female	31/170	18.2	9/284	3.2	6.8	3.2–14.7	<0.001
Smoking status							
Smoking	21/123	17.1	9/402	2.2	9.0	4.0–20.2	<0.001
Non-smoking	29/165	17.6	35/461	7.6	2.6	1.5–4.4	<0.001
Histological subtypes							
Adenocarcinoma§	60/300	20.0	23/628	3.7	6.6	4.0–10.9	<0.001
Non-adenocarcinoma	6/57	10.5	5/175	2.9	4.0	1.2–13.6	0.027
Cancer stage							
I–IIIA	24/148	16.2	17/513	3.3	5.2	2.9–10.8	<0.001
IIIB–IV	42/208	20.2	11/291	3.8	6.4	3.2–12.9	<0.001

§Adenocarcinoma includes adenocarcinoma and adenosquamous carcinoma.
Abbreviations: OR, odds ratio; CI, confidence interval. Yrs, years old.

relatively older age. While chromosomal abnormalities may result in structural changes of critical proteins and hence more aggressive tumors which require less time to become overt diseases. In favor of this point, other cancer types including anaplastic large cell lymphomas, inflammatory myofibroblastic tumor, and neuroblastoma that harbor *ALK* rearrangements occur predominantly in children and young adults. Other fusion genes in NSCLC, such as *ROS1* and *RET* are also associated with younger age at diagnosis[30,31], suggesting a class-specific characteristic of fusion genes that differ from point mutations or indels (insertions and deletions) such as Kras mutations and PI3CA mutations[32,33]. Therefore, in very young NSCLC patients, the tests of *ALK* rearrangements should be given priority over *EGFR* mutations (especially in those younger than 30 years old). Furthermore, in vitro and in vivo studies to investigate the biological difference between these two oncogenic mechanisms are warranted.

Due to the remarkable clinical benefits of tyrosine-kinase inhibitors in NSCLC patients who harbor corresponding driver mutations, pretreatment multiplex genetic tests should be performed to guide therapeutic decisions. However, when tumor tissue is scarce, we may use clinicopathological features to predict specific genetic aberrations. Our study for the first time demonstrates that age at diagnosis alone can be a valuable tool to predict the presence of *EML4-ALK* rearrangement with fair sensitivity and specificity. While smoking pack-years but not age at diagnosis may predict *EGFR* mutations, as supported by previous and current study. However, the predictive power of smoking-pack-years reported here is less evident and we suggest not omitting heavier smokers from *EGFR* mutation testing in East Asian population.

Our study have several limitations. First, this is a single-institution study. However, we prospectively enrolled consecutive NSCLC
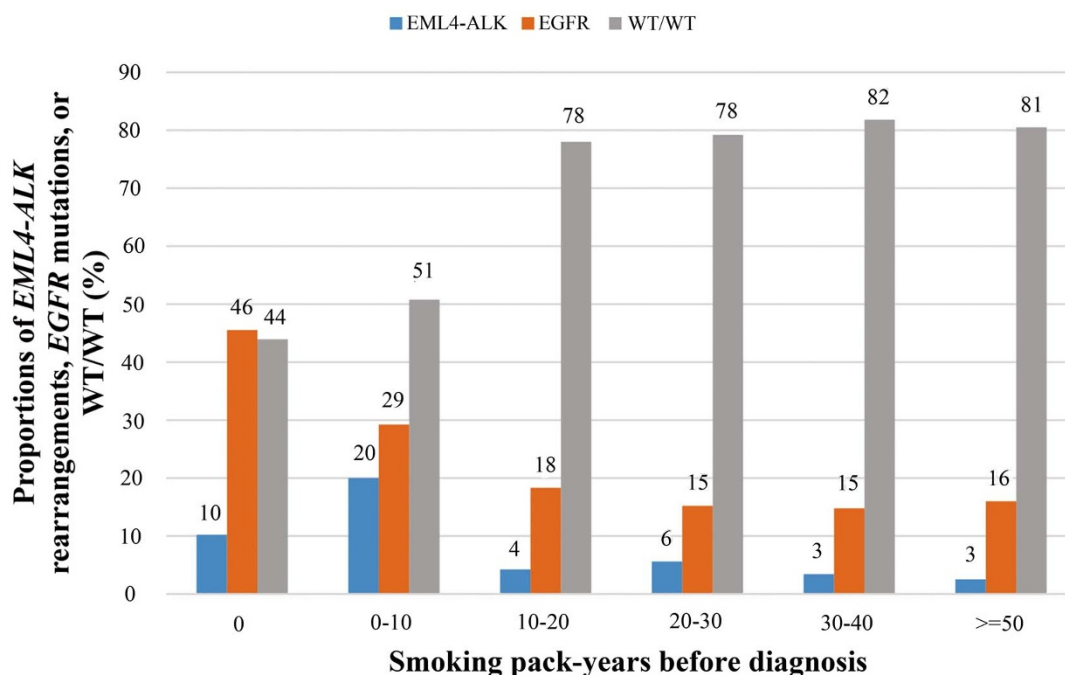

Figure 5 | The incidence of *EML4-ALK* rearrangements, *EGFR* mutations, and *WT/WT* in non-small-cell lung cancer patients according to total smoking pack-years before diagnosis. *WT/WT*, wild type *ALK* and *EGFR*.
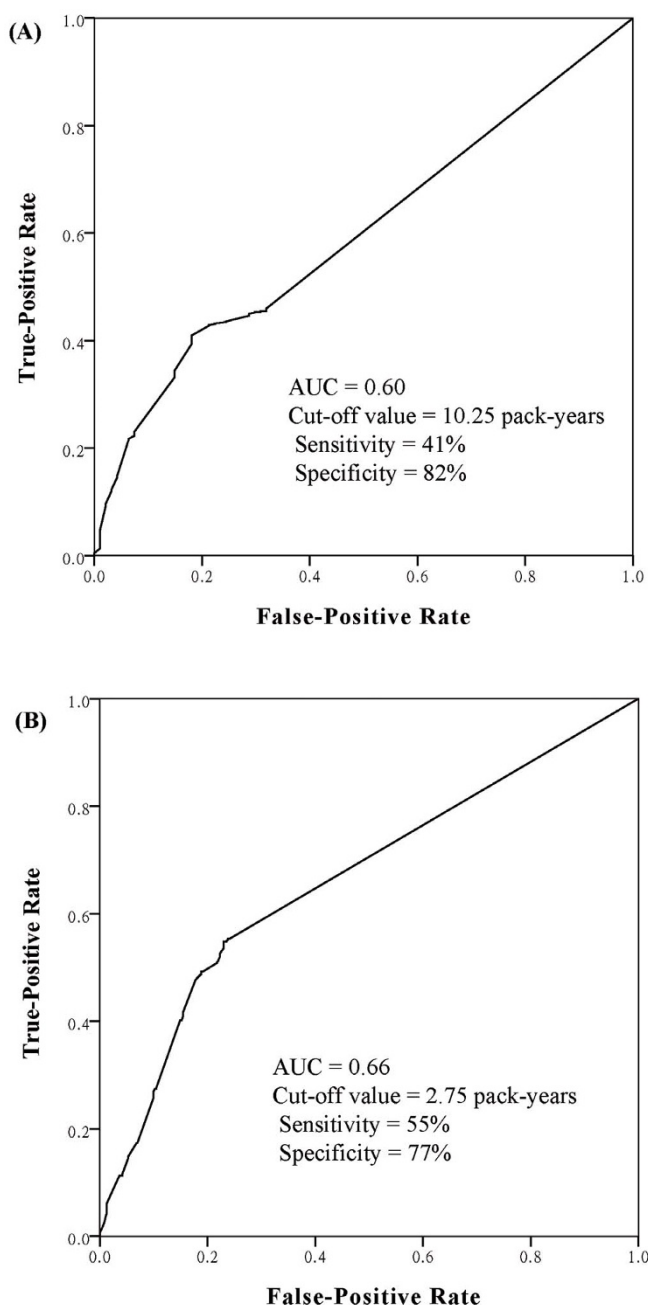
Figure 6 | Receiver operating characteristics (ROC) curves for total smoking pack-years before diagnosis as predictors of (A) *EML4-ALK* rearrangements and (B) *EGFR* mutations in non-small-cell lung cancer. AUC, areas under ROC curve.

etc. have also been reported in previous studies[35]. Whether these driver share similar phenomenon needs further investigation.

In summary, we show that age at diagnosis alone is a valuable predictor of *EML4-ALK* rearrangements but poorly predicts *EGFR* mutations in NSCLC. Smoking pack-years may predict *EGFR* mutations though with limited power. We recommend the detection of *EGFR* mutations should not be confined to patients with "advantageous" features like younger patients, females and non-smoking. The results should help assess the likelihood of these two genetic aberrations based on available clinicopathological features and understand the biological implications of different driver mutations.

Methods

Patients and sample collection. This cross-sectional study was to determine the overall incidence of *EML4-ALK* rearrangement and *EGFR* mutation in Chinese Han population diagnosed as NSCLC; to investigate the distinct clinicopathological pathological features of patients who harbored *EML4-ALK* rearrangements or *EGFR* mutations; and to evaluate the predictive value of age at diagnosis and smoking pack-years for these two genetic aberrations. Patients who met the following criteria were prospectively enrolled: histologically or cytologically proven NSCLC patients by two independent pathologists (Y. Li and J.T. Jin); aged 18 years or older; able to provide informed consent; available and sufficient tumor tissue (biopsy or surgical specimen) for genomic analysis. Specimens were obtained from two sources: fresh-frozen tumor samples from the Biobank of SYSUCC and formalin-fixed, paraffin-embedded (FFPE) tissue submitted to the Department of Pathology (within 5 years before enrollment). The study was conducted in accordance with the Declaration of Helsinki, International Conference on Harmonisation Guidelines for Good Clinical Practice and was approved by the Ethics Committee of SYSUCC. Informed consent was obtained for each participant before the acquisition of tumor tissue.

Genetic analysis. *EGFR* mutations were detected using PCR-based direct sequencing of exons 18–21 as previously described[31]. Briefly, genomic DNA was extracted from either tumors embedded in paraffin blocks or from fresh frozen tumors. PCR amplification was done using HotStarTaq DNA polymerase (Qiagen Inc., Valencia, CA) with a forward primer (5′-GGATCGGCCTCTTCATGC3′) and a reverse primer (5′-TAAAATTGATTCCAATGCCATCC-3′). PCR products were sequenced directly using Applied Biosystems PRISM dye terminator cycle sequencing method (Perkin-Elmer Corp., Foster City, CA) with ABI PRISM 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). Any in-frame deletions in exon 19 or point mutations in exon 21 (L858R or L861Q substitutions), which confer sensitivity to EGFR-TKIs therapy, were considered as *EGFR* mutant. *EML4-ALK* rearrangements were detected by means of Fluorescence in situ Hybridization (FISH) using a break-apart probe to the ALK gene (Vysis LSI ALK Dual Color, Break Apart Rearrangement Probe; Abbott Molecular) per manufacturer's instructions. At least 100 representative tumor cells were counted. The results obtained by FISH were analyzed using an Olympus fluorescence microscope equipped with orange, green, and 4′, 6-diamidino-2-phenylindole filters. Images were captured using the Video Test Image Analysis System. FISH-positive cases were defined as ≥15% of the tumor cells that showed a split red and green signal and/or an isolated (single) red signal. Otherwise, the specimen was classified as ALK FISH negative.

Clinicopathological data. Clinicopathological features including age at diagnosis, gender, smoking history, pathological types, differentiation, cancer stage, symptoms at first onset and family history of malignant tumors were carefully collected. Lung cancer was histologically classified as adenocarcinoma, squamous cell carcinoma, adenosquamous carcinoma and other subtypes. Cancer stage was determined according to TNM classification according to the Union for International Cancer Control and the American Joint Committee on Cancer staging system, 7th edition[32]. Smokers were defined as those who had more than 100 lifetime cigarettes. Smoking pack-years was calculated as average number of cigarettes per day/20× years smoking and was treated as a continuous variable.

Statistical analysis. Chi-square test (or Fisher exact test) and independent-samples t-test were applied to explore the univariate association between clinicopathological variables and specific genetic aberrations, for categorical and continuous data, respectively. All variables that were univariately associated at the level of α < 0.2 were included in the multivariate logistic regression model. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated to denote the association. The receiver operating characteristic (ROC) curve methodology was applied to assess the ability of age at diagnosis or smoking pack-years to predict *EML4-ALK* rearrangements and *EGFR* mutations. The diagnostic accuracy was presented as the area under the ROC curve (AUC). All statistical calculations were performed using SPSS version 21.0 (SPSS, Inc., Chicago, IL.) A two-tail P value of < 0.05 was judged significant.

patients seen in our hospital. We believe these unselected patients are fair representative of NSCLC patients across different pathological types. Second, we did not carry out survival analysis due to immature survival data. It would be interesting to evaluate the prognostic value of clinicopathological variables and mutation types. A recent study by Li C et al. shows that no survival difference was noted in lung adenocarcinoma according to different driver mutations[34]. Yet, some studies found that ALK+ was associated with worse disease-free survival in NSCLC[35,36]. Overall, the prognostic value of different driver mutations are controversial which is probably due to more complicated treatment options in the era of targeted therapy. Third, we only focused on two currently druggable targets in NSCLC. Other driver mutations including KRAS, BRAF, HER2, MET, PTEN, RET,

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA Cancer J Clin.* **64**, 9–29 (2014).
2. Mitsudomi, T. *et al.* Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor

receptor (WJTOG3405): an open label, randomised phase 3 trial. *Lancet Oncol.* **11**, 121–128 (2010).

3. Rosell, R. *et al.* Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **13**, 239–246 (2012).

4. Wu, Y. L. *et al.* Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. *Lancet Oncol.* **15**, 213–222 (2014).

5. Cross, D. A. *et al.* AZD9291, an Irreversible EGFR TKI, Overcomes T790M-Mediated Resistance to EGFR Inhibitors in Lung Cancer. *Cancer Discov.* **4**, 1046–1061 (2014).

6. Shaw, A. T. *et al.* Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med.* **368**, 2385–2394 (2013).

7. Shaw, A. T. *et al.* Ceritinib in ALK-rearranged non-small-cell lung cancer. *N Engl J Med.* **370**, 1189–1197 (2014).

8. Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med.* **350**, 2129–2139 (2004).

9. Pao, W., Iafrate, A. J. & Su, Z. Genetically informed lung cancer medicine. *J Pathol.* **223**, 230–240 (2011).

10. Shi, Y. *et al.* A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). *J Thorac Oncol.* **9**, 154–162 (2014).

11. Pham, D. *et al.* Use of cigarette-smoking history to estimate the likelihood of mutations in epidermal growth factor receptor gene exons 19 and 21 in lung adenocarcinomas. *J Clin Oncol.* **24**, 1700–1704 (2006).

12. Jida, M. *et al.* Usefulness of cumulative smoking dose for identifying the EGFR mutation and patients with non-small-cell lung cancer for gefitinib treatment. *Cancer Sci.* **100**, 1931–1934 (2009).

13. Choi, Y. H. *et al.* Association between age at diagnosis and the presence of EGFR mutations in female patients with resected non-small cell lung cancer. *J Thorac Oncol.* **5**, 1949–1952 (2010).

14. Zhang, Y. *et al.* Frequency of driver mutations in lung adenocarcinoma from female never-smokers varies with histologic subtypes and age at diagnosis. *Clin Cancer Res.* **18**, 1947–1953 (2012).

15. Ueno, T. *et al.* Impact of age on epidermal growth factor receptor mutation in lung cancer. *Lung Cancer.* **78**, 207–211 (2012).

16. Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).

17. Soda, M. *et al.* A mouse model for EML4-ALK-positive lung cancer. *Proc Natl Acad Sci U S A.* **105**, 19893–19897 (2008).

18. Kwak, E. L. *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med.* **363**, 1693–1703 (2010).

19. Camidge, D. R. *et al.* Activity and safety of crizotinib in patients with ALK-positive non-small-cell lung cancer: updated results from a phase 1 study. *Lancet Oncol.* **13**, 1011–1019 (2012).

20. Zhang, X. *et al.* Fusion of EML4 and ALK is associated with development of lung adenocarcinomas lacking EGFR and KRAS mutations and is correlated with ALK expression. *Mol Cancer.* **9**, 188 (2010).

21. Shaw, A. T. *et al.* Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK. *J Clin Oncol.* **27**, 4247–4253 (2009).

22. Fallet, V. *et al.* Prospective screening for ALK: Clinical features and outcome according to ALK status. *Eur J Cancer.* **50**, 1239–1246 (2014).

23. Rodig, S. J. *et al.* Unique clinicopathologic features characterize ALK-rearranged lung adenocarcinoma in the western population. *Clin Cancer Res.* **15**, 5216–5223 (2009).

24. Koivunen, J. P. *et al.* EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin Cancer Res.* **14**, 4275–4283 (2008).

25. Inamura, K. *et al.* EML4-ALK lung cancers are characterized by rare other mutations, a TTF-1 cell lineage, an acinar histology, and young onset. *Mod Pathol.* **22**, 508–515 (2009).

26. Shinmura, K. *et al.* EML4-ALK fusion transcripts, but no NPM-, TPM3-, CLTC-, ATIC-, or TFG-ALK fusion transcripts, in non-small cell lung carcinomas. *Lung Cancer.* **61**, 163–169 (2008).

27. Girard, N. *et al.* Nomogram to predict the presence of EGFR activating mutation in lung adenocarcinoma. *Eur Respir J.* **39**, 366–372 (2012).

28. Tokumo, M. *et al.* The relationship between epidermal growth factor receptor mutations and clinicopathologic features in non-small cell lung cancers. *Clin Cancer Res.* **11**, 1167–1173 (2005).

29. D'Angelo, S. P. *et al.* Incidence of EGFR exon 19 deletions and L858R in tumor specimens from men and cigarette smokers with lung adenocarcinomas. *J Clin Oncol.* **29**, 2066–2070 (2011).

30. Bergethon, K. *et al.* ROS1 rearrangements define a unique molecular class of lung cancers. *J Clin Oncol.* **30**, 863–870 (2012).

31. Wang, R. *et al.* RET fusions define a unique molecular and clinicopathologic subtype of non-small-cell lung cancer. *J Clin Oncol.* **30**, 4352–4359 (2012).

32. Riely, G. J. *et al.* Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin Cancer Res.* **14**, 5731–5734 (2008).

33. Kawano, O. *et al.* PIK3CA mutation status in Japanese lung cancer patients. *Lung Cancer.* **54**, 209–215 (2006).

34. Li, C. *et al.* Prognostic value analysis of mutational and clinicopathological factors in non-small cell lung cancer. *PLoS One.* **9**, e107276 (2014).

35. Zhou, J. X. *et al.* Oncogenic driver mutations in patients with non-small-cell lung cancer at various clinical stages. *Ann Oncol.* **24**, 1319–1325 (2013).

36. Yang, P. *et al.* Worse disease-free survival in never-smokers with ALK+ lung adenocarcinoma. *J Thorac Oncol.* **7**, 90–97 (2012).

Acknowledgments

Author contributions

S.H. and L.Z. conceived and designed the study; S.H., W.F., Z.H., T.Z. and Y.Y. performed the study; S.H., W.F., Y.T., Y.M., Y.Z. and C.X. analyzed the data; all the authors (S.H., W.F., Z.H., T.Z., Y.Y., T.Q., Y.T., Y.M., Y.Z., C.X., Y.H., H.Z. and L.Z.) wrote the article. All authors (S.H., W.F., Z.H., T.Z., Y.Y., T.Q., Y.T., Y.M., Y.Z., C.X., Y.H., H.Z. and L.Z.) have read and approved the final manuscript.

Additional information