Eur Respir J 2012; 39: 366–372
DOI: 10.1183/09031936.00010111

Nomogram to predict the presence of *EGFR* activating mutation in lung adenocarcinoma

N. Girard*,§§, C.S. Sima#,§§, D.M. Jackman¶, L.V. Sequist+, H. Chen§, J.C-H. Yangf, H. Ji, B. Waltman+, R. Rosell##, M. Taron##, M.F. Zakowski¶¶, M. Ladanyi¶¶, G. Riely++ and W. Pao*,++**

ABSTRACT: Epidermal growth factor receptor (*EGFR*) tumour genotyping is crucial to guide treatment decisions regarding the use of EGFR tyrosine kinase inhibitors in nonsmall cell lung cancer (NSCLC). However, some patients may not be able to obtain tumour testing, either because tissue is limited and/or tests are not routinely offered. Here, we aimed to build a model-based nomogram to allow for prediction of the presence of *EGFR* mutations in NSCLC.

We retrospectively collected clinical and pathological data on 3,006 patients with NSCLC who had their tumours genotyped for *EGFR* mutations at five institutions worldwide. Variables of interest were integrated in a multivariate logistic regression model.

In the 2,392 non-Asian patients with lung adenocarcinomas, the most important predictors of harbouring *EGFR* mutation were: lower tobacco smoking exposure (OR 0.41, 95% CI 0.37–0.46), longer time interval between smoking cessation and diagnosis (OR 2.19, 95% CI 1.71–2.80), advanced stage (OR 1.58, 95% CI 1.18–2.13), and papillary (OR 4.57, 95% CI 3.14–6.66) or bronchioloalveolar (OR 2.84, 95% CI 1.98–4.06) histologically predominant subtype. A nomogram was established and showed excellent discriminating accuracy: the concordance index on an independent validation dataset was 0.84.

As clinical practices transition to incorporating genotyping as part of routine care, this nomogram could be highly useful to predict the presence of *EGFR* mutations in lung adenocarcinoma in non-Asian patients when mutational profiling is not available or possible.

KEYWORDS: Adenocarcinoma, epidermal growth factor receptor mutations, lung cancer, nomogram, prediction score, tyrosine kinase inhibitor

Nonsmall cell lung cancer (NSCLC) is the most frequent cause of cancer-related death worldwide, accounting for >1 million deaths per year [1]. NSCLC is primarily comprised of three different types: squamous cell carcinoma, large cell carcinoma and adenocarcinoma. Adenocarcinoma now comprises >50% of all cases of lung cancer in the USA and Western Europe [2, 3].

Somatic mutations in the epidermal growth factor receptor (*EGFR*) gene are present in a subset of lung adenocarcinomas. Biological data regarding *EGFR* mutations have been extensively reviewed elsewhere [4]. Briefly, *EGFR* mutations primarily occur in the tyrosine kinase domain, mostly involving deletions in exon 19 and a recurrent point mutation (L858R) in exon 21 [5, 6]. These mutations lead to constitutive activation of the receptor, independent of ligand binding, and are associated with increased sensitivity to the specific EGFR tyrosine kinase inhibitors (TKIs) gefitinib and erlotinib [5]. Clinically, response rates to these agents are >70% in tumours harbouring *EGFR* activating mutations [7–9]. Recent randomised phase III trials (Iressa Pan-Asia Study (IPASS), West Japan Thoracic Oncology Group 3405 and North-East Japan Study Group) have found that, for patients with *EGFR* mutant tumours, first-line gefitinib leads to a longer progression-free survival compared with standard platinum-based doublet chemotherapy [7–9]. Thus, genotyping of *EGFR* is crucial to guide treatment decisions regarding the use of EGFR TKIs and is becoming a standard recommendation in the pretreatment work-up of patients with lung adenocarcinoma [10–12].

Unfortunately, in many countries, routine access to mutation testing has been hampered by a lack of well-established molecular diagnostic laboratories

AFFILIATIONS
*Human Oncology and Pathogenesis Program,
#Dept of Epidemiology and Biostatistics,
¶¶Dept of Pathology, and
++Thoracic Oncology Service, Memorial Sloan-Kettering Cancer Center, New York, NY,
¶Medical Oncology, Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, and
+Massachusetts General Hospital Cancer Center, Harvard Medical School, Boston, MA, USA.
§Dept of Thoracic Surgery, The Cancer Hospital of Fudan University, and
**Laboratory of Molecular Cell Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.
fGraduate Institute of Oncology, National Taiwan University College of Medicine, Taipei, Taiwan.
##Catalan Institute of Oncology, Hospital Germans Trias I Pujol, Ctra Canyet, Badalona, Spain.
§§These authors contributed equally to the study.

CORRESPONDENCE
N. Girard, Service de Pneumologie U80, Hôpital Louis Pradel, 28 avenue doyen Lépine, 69677 Lyon, Bron cedex, France
E-mail: nicolas.girard@chu-lyon.fr

Received:
Jan 19 2011
Accepted after revision:
June 16 2011
First published online:
July 20 2011

European Respiratory Journal
Print ISSN 0903-1936
Online ISSN 1399-3003

or by economic and/or regulatory issues. For example, from 2008 to 2009, only 16% and 11% of metastatic adenocarcinoma cases were genotyped for *EGFR* activating mutations in France and Spain, respectively. Moreover, many patients with metastatic NSCLC are diagnosed with needle biopsies, limiting the amount of tissue available for mutation testing. In IPASS, tumour specimens were collected for only 66% of patients and *EGFR* mutation status could ultimately be assessed in only 64% of those cases [7]. In these settings, the value of an accurate prediction tool for the presence of *EGFR* mutations in patients' tumours, based upon clinical characteristics, is indisputable.

Statistical prediction models are widely used for cancer outcome prediction. Among those, nomograms create user-friendly graphical representations of complex models to generate the probability of an event based on the individual profile of each patient [13]. Here, we integrated multiple clinical and pathological factors from a large, well-annotated international cohort of genotyped lung tumours to develop a multivariate logistic regression model and an associated nomogram to predict the presence of *EGFR* activating mutations.

PATIENTS AND METHODS

Patients

We retrospectively collected clinical and pathological data on 3,006 patients with NSCLC, including 2,856 with adenocarcinomas, who had their tumours genotyped for *EGFR* activating mutations at five institutions worldwide: 1,990 adenocarcinoma patients were from Memorial Sloan-Kettering Cancer Center (MSKCC) (New York, NY, USA); 297 were from the Dana-Farber Cancer Institute (DFCI) (Boston, MA, USA); 269 were from Massachusetts General Hospital Cancer Center (MGH) (Boston, MA); 213 were from the Cancer Hospital of Fudan University (CHFU) (Shanghai, China); and 87 were from the National Taiwan University Hospital (NTUH) (Taipei, Taiwan) [14]. The Catalan Institute of Oncology (CIO) (Hospital Germans Trias I Pujol, Badalona, Spain) also contributed data on 217 patients (198 adenocarcinomas) who were ultimately not included in the analyses because of missing data on key variables, such as smoking history.

Starting in 2004 at NTUH, DFCI, MGH and MSKCC, 2005 at CIO, and 2007 at CHFU, *EGFR* genotyping was systematically and prospectively performed for all consecutive, cytologically or histologically proven adenocarcinoma tumours with available material for testing, without any selection based on additional clinical or pathological criteria. Therefore, patients included in our analysis are representative of all patients treated at the five institutions included.

Data were collected in July 2009 under local institutional review board (IRB)-approved protocols. The present study was approved by each institution IRB.

Mutational analyses

For this analysis, we considered only drug-sensitive *EGFR* activating mutations, *i.e.* kinase domain mutations associated with sensitivity to EGFR TKIs: exon 19 deletions, exon 21 mutation (*i.e.* L858R) and other point mutations (G719A/C/S and L861Q). Tumours harbouring other *EGFR* mutations were excluded. Direct sequencing was the standard method used for *EGFR* genotyping (exons 18–21) at DFCI, MGH, CIO, CHFU and NTUH. At MSKCC, *EGFR* mutational analysis was performed both using direct sequencing and PCR-based, mutation-specific assays to detect the two most common *EGFR* mutations, *i.e.* exon 19 deletions and L858R point mutations [15]. We also retrieved the *KRAS* mutational status of NSCLC tumours, if available. *KRAS* mutations were assessed using direct sequencing [16].

Clinical and pathological data

Clinical and pathological variables of interest included: age, sex, race/ethnicity, histological type (adenocarcinoma, squamous cell carcinoma or large cell carcinoma), and for adenocarcinoma tumours, histologically predominant subtype [17], tumour stage according to the 6th American Joint Committee on Cancer TNM (tumour, metastasis and node) system that was standard at time the data were generated [18], and smoking variables (smoking status, smoking quantity (measured in pack-yrs) for smokers and time interval since quitting for former smokers). Patients were categorised as never-smokers (<100 lifetime cigarettes), former smokers (quit >1 yr prior to diagnosis) or current smokers (continued smoking within 1 yr of diagnosis). Pathological review was performed at each institution. We also retrieved results from routine immunohistochemical studies performed with an antibody against thyroid transcription factor (TTF)-1, a marker of terminal respiratory unit adenocarcinoma, which is more likely to harbour *EGFR* mutation [19]. However, because data were lacking, we could not include this covariate in the final model.

Statistical analysis

Analyses were performed on patients diagnosed with NSCLC adenocarcinoma, separately for non-Asian and Asian patients, resulting in one model for each of the two populations. The Chi-squared and two-sample paired t-tests were used to compare the univariate association between clinical parameters and the presence of *EGFR* or *KRAS* mutations, for categorical and continuous variables, respectively. The methodology used to develop the prediction nomogram was previously described [13, 20]. Briefly, for each model developed, we randomly divided the available data from patients diagnosed with lung adenocarcinoma into a training set (including two-thirds of the patients), used to develop a prediction model, and a validation set (including the remaining one-third). A model-building exercise was performed to develop multivariate logistic regression models that used available clinical and pathological factors to predict the likelihood of an *EGFR* mutation. All variables that were univariately associated with *EGFR* status at a level of $\alpha<0.2$ were candidates for inclusion in the multivariate model, and were retained in the model if they remained significantly associated with the outcome. Those variables with a markedly skewed distribution were log-transformed; the linearity assumption was further checked for continuous and ordinal variables and, if the functional form of the relationship with the outcome suggested, they were modelled using quadratic polynomials. Levels of the categorical variables were combined if they had similar effects. Interactions between all combinations of two variables were tested but were not included in the final model because their effect did not reach a level of significance of $\alpha<0.05$ and their addition to the model did not improve the model's predictive accuracy by >1%.

The discriminatory ability of the model was quantified using the concordance index, which is numerically equivalent to the area

under the receiver operating characteristic curve. Concordance index indicates the probability that, in a randomly selected pair in which one patient has and the other patient does not have *EGFR* mutation, the model correctly predicts a higher probability of mutation for the first patient. In practice, concordance index ranges from 0.5 (no discriminatory ability, equivalent to a coin toss) to 1.0 (perfect discrimination) [21].

Calibration was assessed by plotting the probabilities predicted by the logistic model *versus* the actual probability. Since the individual actual probability is either 1 or 0 (for patients with and without *EGFR* mutation, respectively), calibration plots were built using local regression, a method commonly used for data smoothing that relies on approximating each data point with a low-degree polynomial using the neighbouring values. A plot along the 45° line will correspond to a model in which the predicted probabilities are identical with the actual probabilities, therefore indicating perfect calibration. Both discrimination and calibration of the prediction models were calculated on the independent validation dataset.

To protect against the influence of the initial random split, the analysis was repeated in 10 randomly split training and validation datasets, using 10-fold cross-validation, refitting the model to omit one-tenth of the observations to obtain prediction for the omitted tenth. Statistical calculations were performed using SAS version 9.2 (SAS Institute Inc., Cary, NC, USA), and the Design and locfit packages in R (The R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

EGFR *mutations in lung adenocarcinomas of non-Asian patients*

Population characteristics

We focused our analysis on the largest population in our cohort, namely non-Asian patients (including white and black, and Hispanic and non-Hispanic), with lung adenocarcinoma. To maximise the chances of building an effective model, we included only cases for which all smoking variables were available (n=2,392). 1,579 (66%) patients were randomly included in a training dataset and the remaining 813 (34%) patients formed a validation dataset. Demographic, clinical and pathological characteristics of the patients are presented in table 1. Overall, 604 (25%) tumours harboured an *EGFR* activating mutation, corresponding to exon 19 deletion in 54% of cases and L858R substitution in 37% of cases.

Patients with *EGFR* mutant tumours were more likely to be females (72% *versus* 62% of patients; p<0.001) and to be diagnosed with advanced-stage disease (55% *versus* 45%; p<0.001) compared with the patients with *EGFR* wild-type tumours. There was no significant difference between the two groups with respect to age (64±12 *versus* 65±11 yrs; p=0.370) or race/ethnicity (94% of patients were white in both groups). 55% of the patients with *EGFR* mutant tumours were never-smokers, compared with 19% of those with *EGFR* wild-type tumours (p<0.001). For former smokers, median time interval since quitting was 27 yrs among *EGFR* mutant and 20 yrs among *EGFR* wild-type patients.

Tumours with *EGFR* mutations were more likely to harbour papillary (30% *versus* 12% of patients; p<0.001) and bronchioloalveolar predominant histological subtypes (25% *versus* 17%; p<0.001), and less likely to exhibit acinar (18% *versus* 39%; p<0.001) or solid architecture (9% *versus* 14%; p=0.002). TTF-1 expression was more frequent in *EGFR* mutant tumours (97% *versus* 84% of tested tumours; p<0.001). Histological subtype was not otherwise specified (NOS) or could not be assessed for 18% of adenocarcinomas, because the diagnosis was made either on small-size tumour tissue samples (22% of NOS cases) or cytological specimens (78% of NOS cases).

TABLE 1 Clinical and pathological characteristics associated with epidermal growth factor receptor (*EGFR*) mutational status in 2,392 non-Asian patients with lung adenocarcinoma

	EGFR wild-type	*EGFR* mutant	p-value
Patients n	1788	604	
Population characteristics			
Age yrs	65±11	64±12	0.370
Sex			
Males	682 (38)	169 (28)	<0.001
Females	1106 (62)	435 (72)	
Race			
White	1688 (94)	570 (94)	0.970
Other	100 (6)	34 (6)	
Smoking status			
Never-smokers	333 (19)	330 (55)	<0.001
Smokers	1455 (81)	274 (45)	
Total tobacco quantity pack-yrs	40±28	15±17	<0.001
Former smokers	1085 (60)	251 (42)	
Time since quitting yrs	20±14	27±14	<0.001
Current smokers	370 (21)	23 (4)	
Tumour characteristics			
Histology			
Predominant subtype			
Acinar	700 (39)	110 (18)	<0.001
Papillary	209 (12)	181 (30)	<0.001
BAC	312 (17)	148 (25)	<0.001
Solid	252 (14)	55 (9)	0.002
NOS	315 (18)	110 (18)	0.741
TTF-1 expression[#]			
Yes	764 (43)	284 (47)	<0.001
No	149 (8)	8 (2)	
Stage			
I–IIIA	984 (55)	274 (45)	<0.001
IIIB–IV	804 (45)	330 (55)	
KRAS mutational status[¶]			
Mutant	330 (18)	0 (0)	<0.001
Wild-type	1172 (66)	507 (84)	

Data are presented as median±SD or n (%), unless otherwise stated. BAC: bronchioloalveolar carcinoma; NOS: not otherwise specified; TTF: thyroid transcription factor. [#]: data missing for 875 (49%) *EGFR* wild-type tumours and 312 (52%) *EGFR* mutant tumours; [¶]: data missing for 286 (16%) *EGFR* wild-type tumours and 97 (16%) *EGFR* mutant tumours.

Prediction nomogram

The results of a multivariate model predicting the probability of *EGFR* mutation in non-Asian patients with lung adenocarcinoma

are presented in table 2. The most important predictors of harbouring *EGFR* activating mutation were: smoking history, measured by lower total tobacco smoking exposure (log-transformed pack-yrs; OR 0.41, 95% CI 0.37–0.46) and longer time interval between tobacco smoking cessation and NSCLC diagnosis (OR 2.19, 95% CI 1.71–2.80); advanced stage at diagnosis (OR 1.58, 95% CI 1.18–2.13), and papillary (OR 4.57, 95% CI 3.14–6.66) or bronchioloalveolar (OR 2.84, 95% CI 1.98–4.06) histologically predominant subtype. These predictors were not different for L858R mutations and exon 19 deletions.

Based upon this analysis, we generated a nomogram to predict the presence of activating *EGFR* mutation in tumours (fig. 1). The concordance index calculated for the independent validation dataset was 0.84 (95% CI 0.80–0.86), indicating excellent accuracy in discriminating *EGFR* mutant *versus EGFR* wild-type cases in this group of patients. In addition, the calibration plot (fig. 2) indicates that the nomogram-predicted probabilities compared very well with actual probabilities. Moreover, additional internal validation indicated no change between the original concordance index and that based on 10-fold cross-validation.

Influence of KRAS *mutations on the nomogram*

We tested the effect of *KRAS* mutational status on the nomogram. *KRAS* mutations are mutually exclusive with *EGFR* mutations in lung adenocarcinoma and *KRAS* mutant tumours represent a subset of *EGFR* wild-type tumours that do not respond to EGFR TKIs [16, 22]. Given the higher frequency of *KRAS* mutations in NSCLC tumours from non-Asian patients and technical issues that may make *KRAS* genotyping easier, *KRAS* mutations have been used by some institutions as a negative surrogate marker for the presence of *EGFR* mutations. *KRAS* mutational status was available for 82% of cases; 17% of genotyped tumours harboured *KRAS* mutations. In our cohort, we did not identify clinical or pathological variables that were significantly associated with the presence of *KRAS* mutations. A multivariate model restricted to

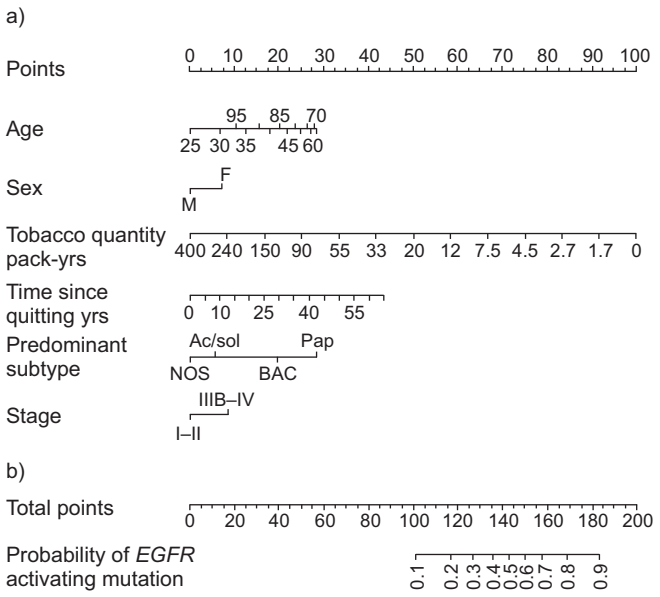

FIGURE 1. Nomogram including clinical and pathological variables to predict the presence of epidermal growth factor receptor (*EGFR*) activating mutation in adenocarcinomas from non-Asian patients. a) Locate the total pack-year quantity that was smoked by the patient on the tobacco quantity axis. Draw a line straight upwards to the points axis to determine how many points towards *EGFR* mutation the patient receives for tobacco consumption. Repeat the process for the other axes (age, sex, time since quitting smoking, predominant histological subtype and stage), each time drawing a straight line upwards to the points axis. b) Sum the points achieved for each predictor and locate this sum on the total points axis. Draw a line straight down to the bottom line to find the probability of *EGFR* mutation in the tumour. M: male; F: female; NOS: not otherwise specified; Ac/sol: acinar or solid; BAC: bronchioloalveolar carcinoma; Pap: papillary.

TABLE 2 Multivariable logistic regression analysis of clinical and pathological variables predicting the presence of epidermal growth factor receptor (*EGFR*) activating mutation in lung adenocarcinomas from non-Asian patients

	OR (95% CI)	p-value
Age[#] per 5 yrs	0.89 (0.80–1.00)	0.007
Female *versus* male	1.47 (1.10–1.97)	0.009
Natural log-transformed total pack-yrs	0.41 (0.37–0.46)	<0.001
Time since quitting[¶] per yr	2.19 (1.71–2.80)	<0.001
Stage IIIB–IV *versus* I–IIIA	1.58 (1.18–2.13)	0.003
Predominant subtype		
Papillary *versus* acinar/solid	4.57 (3.14–6.66)	<0.001
BAC *versus* acinar/solid	2.84 (1.98–4.06)	<0.001
NOS *versus* acinar/solid	1.36 (0.88–2.11)	0.170

BAC: bronchioloalveolar carcinoma; NOS: not otherwise specified. [#]: included in the model as a quadratic term; odds ratio corresponds to an increase in the probability of *EGFR* mutation associated with a change in age from 55 to 60 yrs. [¶]: variable applied to former smokers only.

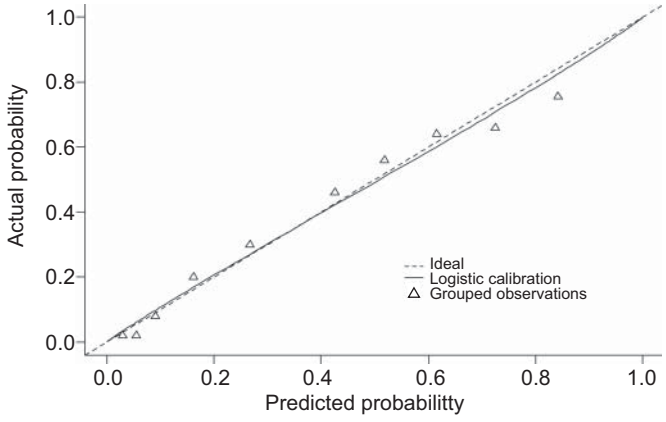

FIGURE 2. Calibration graph for the validation dataset. Calibration was assessed by plotting the predicted *versus* actual probability of harbouring epidermal growth factor receptor (*EGFR*) mutation in lung adenocarcinoma tumours in the validation cohort of 813 non-Asian patients. Predicted probabilities of *EGFR* mutation for are plotted on the *x*-axis. Observed frequencies of *EGFR* mutation are plotted on the *y*-axis. The dashed line indicates reference line of an ideal nomogram. The solid line indicates the actual performance of the nomogram. The concordance index was 0.84.

patients who had *KRAS* wild-type tumours yielded very similar results to the model described above. The prediction accuracy of this model on the independent validation dataset was 0.83.

EGFR *mutations in lung adenocarcinomas of East Asian patients*

We analysed separately a group of 464 East Asian patients with lung adenocarcinoma for whom all smoking covariates were available (table 3). Variables that were significantly associated with the presence of activating *EGFR* mutation in the tumour were: female sex (66% *versus* 46% of patients; p<0.001), never-smoker status (79% *versus* 57%; p<0.001) and lower tobacco quantity for former/current smokers (24±21 *versus* 39±24 pack-yrs; p<0.001). Similar to what was observed in the group of non-Asian patients, *EGFR* mutant tumours were more likely to harbour papillary (20% *versus* 8% of patients; p<0.001) or bronchioloalveolar (22% *versus* 12%; p=0.01) histologically predominant subtypes, and less likely to harbour acinar (43% *versus* 54%; p=0.02) or solid (1% *versus* 7%; p=0.010) features.

TABLE 3 Clinical and pathological characteristics associated with epidermal growth factor receptor (*EGFR*) mutational status in 464 Asian patients with adenocarcinomas

	EGFR wild-type	*EGFR* mutant	p-value
Patients n	179	285	
Population characteristics			
Age yrs	61±12	60±11	0.510
Sex			
Males	97 (54)	97 (34)	<0.001
Females	82 (46)	188 (66)	
Smoking status			
Never-smokers	102 (57)	226 (79)	<0.001
Smokers	77 (43)	59 (21)	
Tobacco quantity pack-yrs	39±24	24±21	<0.001
Former smokers	24 (13)	30 (11)	
Time since quitting yrs	13±12	16±12	0.420
Current smokers	53 (30)	29 (10)	
Tumour characteristics			
Histology			
Predominant subtype			
Acinar	97 (54)	124 (43)	0.020
Papillary	15 (8)	57 (20)	<0.001
BAC	22 (12)	62 (22)	0.010
Solid	12 (7)	4 (1)	0.010
NOS	33 (18)	37 (13)	0.110
Stage			
I–IIIA	95 (53)	174 (61)	0.090
IIIB–IV	84 (47)	111 (39)	
KRAS mutational status[#]			
Mutant	15 (8)	0 (0)	<0.001
Wild-type	124 (69)	231 (82)	

Data are presented as median±SD or n (%), unless otherwise stated. BAC: bronchioloalveolar carcinoma; NOS: not otherwise specified. [#]: data missing for 40 (22%) *EGFR* wild-type tumours and 54 (18%) *EGFR* mutant tumours.

The results of a multivariate model are presented in table 4. Because of the small number of patients in each subgroup, we combined histologically predominant subtype categories as follows: bronchioloalveolar carcinoma/papillary *versus* acinar/solid/NOS. Lower pack-year smoking history (OR 0.63, 95% CI 0.53–0.75), early stage (OR 2.27, 95% CI 1.32–3.85) and bronchioloalveolar carcinoma/papillary predominant histology (OR 2.73, 95% CI 1.51–4.95) remained significantly associated with a higher probability of *EGFR* mutation. This model had a prediction accuracy of 0.75 on the training dataset; however, its accuracy was unsatisfactory when validated on an independent dataset (concordance index was 0.64, 95% CI 0.55–0.73). Notably, in our dataset, only 29% of Asian patients were smokers, resulting in a lack of power of smoking variables to estimate the risk of mutation, especially in the lower-risk group.

DISCUSSION

Tumour *EGFR* genotype has become an important variable in the evaluation of patients with lung cancer. However, in many countries, barriers such as tissue and test accessibility still exist that prohibit patients from obtaining tumour *EGFR* mutation status. Here, we integrated clinical and pathological factors from thousands of patients from the USA and East Asia, to develop a logistic regression-based model to better predict the presence of *EGFR* mutations in NSCLCs, should mutation testing not be available. The nomogram has excellent discrimination properties in non-Asian patients with a concordance index >80%. To our knowledge, this cohort represents the largest dataset analysed to date for comprehensive clinical characteristics associated with *EGFR* mutations.

Other studies have identified clinical and pathological characteristics associated with the presence of *EGFR* activating mutations in lung cancer [6, 17, 23–27]. Consistent with previous results, we found, through multivariate analysis, that smoking history, including never-smoker status, lower tobacco consumption in smokers and longer smoking discontinuation before NSCLC diagnosis in former smokers, were all associated with the presence of *EGFR* activating mutation [17, 23–27]. We also observed an association between bronchioloalveolar and papillary histological subtypes [5, 17, 27]. Contrary to our analyses, previous studies did not identify advanced tumour stage as a significant

TABLE 4 Multivariable logistic regression analysis of clinical and pathological variables predicting the presence of epidermal growth factor receptor (*EGFR*) activating mutation in lung adenocarcinoma tumours from Asian patients

	OR (95% CI)	p-value
Age[#] per 5 yrs	0.98 (0.87–1.09)	0.68
Natural log-transformed total pack-yrs	0.63 (0.53–0.75)	<0.001
Stage IIIB–IV *versus* I–IIIA	0.44 (0.26–0.76)	0.003
BAC/papillary *versus* acinar/solid/NOS histologically predominant subtype	2.73 (1.51–4.95)	<0.001

BAC: bronchioloalveolar carcinoma; NOS: not otherwise specified. [#]: included in the model as a quadratic term; odds ratio corresponds to an increase in the probability of *EGFR* mutation associated with a change in age from 55 to 60 yrs.

predictor of harbouring *EGFR* mutation. Unidentified selection or technical bias may exist in the datasets we analysed, as *EGFR* mutations are rather known to occur early in the lung carcinogenesis process [28].

Overall, compared with prior reports, this study was the first, to our knowledge, to incorporate multiple clinical variables, including smoking history and histological subtype, as well as sex, stage of disease and age, into a user-friendly predictive nomogram for *EGFR* mutations. The additional variables allow for higher accuracy, and the use of test and validation sets further substantiates the validity of the model.

Our nomogram does have some limitations. First, our analyses focused on adenocarcinoma tumours, limiting our analysis of other histologies. However, the majority of *EGFR* mutations are found in adenocarcinomas [24]. Secondly, the proposed nomogram does not apply to Asian patients, who had very different distributions of the predictive variables (smoking history and stage) and considerably higher prevalence of *EGFR* mutations (61% compared with 25% in non-Asian patients) in our dataset. Probably due to these issues, attempts to build a common nomogram that applies to both non-Asian and Asian patients resulted in a predictive instrument with unsatisfactory predictive accuracy. Thirdly, any prediction instrument inherently incorporates a certain degree of uncertainty [21] and individual predictions remain imperfect. Our data come from large academic centres, and the patient population in these institutions may be different from that observed in other medical settings; in general, caution should be exercised if the nomogram is used to predict the likelihood of *EGFR* mutation for patients coming from populations with a different case-mix and a different prevalence of mutation than those studied here.

In IPASS, inclusion criteria were based on clinical characteristics associated with the presence of *EGFR* mutation in the tumour, including never- or light smoker status and adenocarcinoma histology [7]. In the whole intent-to-treat population, for whom subsequent genotyping showed a 59.7% rate of *EGFR* mutation in tumours, progression-free survival was longer in the gefitinib arm than in the chemotherapy arm (hazard ratio 0.74, 95% CI 0.65–0.85; $p<0.0001$) [7]. Based on these data, the threshold for considering the chance of *EGFR* mutation, as predicted by our nomogram, to be clinically significant for recommending EGFR TKI, may be 60%. The use of EGFR TKIs in patients with a moderate probability of a *EGFR* mutant tumour based on the nomogram may not be recommended, as EGFR TKI treatment resulted in a poorer outcome in *EGFR* wild-type tumours compared with standard chemotherapy in the reported trials [7, 29]. Finally, prospective validation of independent datasets will have to be conducted; inclusion of additional Asian patients and non-adenocarcinoma tumours, as well as response to EGFR TKI treatment, would increase the clinical significance of our results.

Ultimately, despite the statistically accurate value of our nomogram to predict the presence of *EGFR* activating mutation in lung adenocarcinoma from non-Asian patients, *EGFR* tumour genotyping should be obtained when possible. Trials that have included ''clinically enriched'' populations, *i.e.* patients more likely to present with *EGFR* mutant tumours, did not achieve improvements in outcome in the intent-to-treat population [7, 30], while trials that studied patients with known *EGFR* mutant tumours did [8, 9]. However, tumour genotyping is not and will not be feasible in all cases. In such instances, our nomogram can then be used to prioritise the use of EGFR TKIs in patients with lung adenocarcinoma.

STATEMENT OF INTEREST

Statements of interest for H. Ji, M. Ladanyi and G. Riely can be found at www.erj.ersjournals.com/site/misc/statements.xhtml

ACKNOWLEDGEMENTS

We thank M.G. Kris (Memorial Sloan-Kettering Cancer Center, New York, NY, USA) for helpful comments, M. Butaney (Dana-Farber Cancer Institute, Boston, MA, USA) for DFCI data assembly and Y.L. Lin (National Taiwan University Hospital, Taipei, Taiwan) for NTUH data retrieval.

REFERENCES

1 American Cancer Society. Cancer Facts and Figures 2007. Atlanta, American Cancer Society, 2007.
2 Travis WB. WHO histological classification of tumours of the lung. *In:* Travis WB, Brambilla A, Muller-Hermelinck HK, *et al*., eds. World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. Lyon, IARC Press, 2004; p 10.
3 Gabrielson E. Worldwide trends in lung cancer pathology. *Respirology* 2006; 11: 533–538.
4 Riely GJ, Politi KA, Miller VA, *et al.* Update on epidermal growth factor receptor mutations in non-small cell lung cancer. *Clin Cancer Res* 2006; 12: 7232–7241.
5 Pao W, Miller V, Zakowski M, *et al.* EGF receptor gene mutations are common in lung cancers from ''never smokers'' and are associated with sensitivity of tumours to gefitinib and erlotinib. *Proc Natl Acad Sci USA* 2004; 101: 13306–13311.
6 Shigematsu H, Gazdar AF. Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers. *Int J Cancer* 2006; 118: 257–262.
7 Mok TS, Wu YL, Thongprasert S, *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009; 361: 947–957.
8 Mitsudomi T, Morita S, Yatabe Y, *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* 2010; 11: 121–128.
9 Maemondo M, Inoue A, Kobayashi K, *et al.* Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* 2010; 362: 2380–2388.
10 Annex I: Summary of Product Characteristics. www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/001016/WC500036358.pdf Date last accessed: January 17, 2011. Date last updated: November 7, 2011.
11 Azzoli CG, Baker S Jr, Temin S, *et al.* American Society of Clinical Oncology Clinical Practice Guideline update on chemotherapy for stage IV non-small-cell lung cancer. *J Clin Oncol* 2009; 27: 6251–6266.
12 National Comprehensive Cancer Network®. www.nccn.org/professionals/physician_gls/PDF/nscl.pdf Date last accessed: January 17, 2011. Date last updated: November 7, 2011.
13 Iasonos A, Schrag D, Raj GV, *et al.* How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol* 2010; 26: 1364–1403.
14 Yang CH, Yu CJ, Shih JY, *et al.* Specific *EGFR* mutations predict treatment outcome of stage III/IV chemonaive NSCLC patients receiving first-line gefitinib monotherapy. *J Clin Oncol* 2008; 26: 2745–2753.

15 Zakowski MF, Ladanyi M, Rekhtman N, *et al.* Reflex testing of lung adenocarcinomas for *EGFR* and *KRAS* mutations: the Memorial Sloan-Kettering experience. *J Clin Oncol* 2008; 26: 22031.
16 Ding L, Getz G, Wheeler DA, *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008; 455: 1069–1075.
17 Motoi N, Szoke J, Riely GJ, *et al.* Lung adenocarcinoma: modification of the 2004 WHO mixed subtype to include the major histologic subtype suggests correlations between papillary and micropapillary adenocarcinoma subtypes, EGFR mutations and gene expression analysis. *Am J Surg Pathol* 2008; 32: 810–827.
18 Mountain CF. Revisions in the International System for Staging Lung Cancer. *Chest* 1997; 111: 1710–1717.
19 Yatabe Y, Kosaka T, Takahashi T, *et al. EGFR* mutation is specific for terminal respiratory unit type adenocarcinoma. *Am J Surg Pathol* 2005; 29: 633–639.
20 Kattan MW, Eastham JA, Stapleton AM, *et al.* A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst* 1998; 90: 766–771.
21 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361–387.
22 Pao W, Girard N. New driver mutations in non-small cell lung cancer. *Lancet Oncol* 2011; 12: 175–180.
23 Pham D, Kris MG, Riely GJ, *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* 2006; 24: 1700–1704.
24 Rosell R, Moran T, Queralt C, *et al.* Screening for epidermal growth factor receptor mutations in lung cancer. *N Engl J Med* 2009; 361: 958–967.
25 Tanaka T, Matsuoka M, Sutani A, *et al.* Frequency of and variables associated with the EGFR mutation and its subtypes. *Int J Cancer* 2010; 126: 651–655.
26 Toyooka S, Matsuo K, Shigematsu H, *et al.* The impact of sex and smoking status on the mutational spectrum of epidermal growth factor receptor gene in non small cell lung cancer. *Clin Cancer Res* 2007; 13: 5763–5768.
27 Dacic S, Shuai Y, Yousem S, *et al.* Clinicopathological predictors of *EGFR/KRAS* mutational status in primary lung adenocarcinomas. *Mod Pathol* 2010; 23: 159–168.
28 Sakamoto H, Shimizu J, Horio Y, *et al.* Disproportionate representation of *KRAS* gene mutation in atypical adenomatous hyperplasia, but even distribution of *EGFR* gene mutation from preinvasive to invasive adenocarcinomas. *J Pathol* 2007; 212: 287–294.
29 Gridelli C, Ciardiello F, Feld R, *et al.* International multicenter randomized phase III study of first-line erlotinib (E) followed by second-line cisplatin plus gemcitabine (CG) *versus* first-line CG followed by second-line E in advanced non-small cell lung cancer (aNSCLC): the TORCH trial. *J Clin Oncol* 2010; 28: 7508.
30 Lee JS, Park K, Kim SW, *et al.* A randomized phase III study of gefitinib (IRESSA^{TM}) *versus* standard chemotherapy (gemcitabine plus cisplatin) as a first-line treatment for never-smokers with advanced or metastatic adenocarcinoma of the lung. *J Thoracic Oncol* 2009; 4: Suppl. 1, S283–S284.