Composite anatomical–clinical–molecular prognostic model in nonsmall cell lung cancer

A. López-Encuentra, F. López-Ríos, E. Conde, R. García-Luján, A. Suárez-Gauthier, N. Mañes, G. Renedo, J.L. Duque-Medina, E. García-Lagarto, R. Rami-Porta, G. González-Pont, J. Astudillo-Pombo, J.L. Maté-Sanz, J. Freixinet, T. Romero-Saavedra, M. Sánchez-Céspedes and A. Gómez de la Camara on behalf of the Bronchogenic Carcinoma Cooperative Group of the Spanish Society of Pneumology and Thoracic Surgery (GCCB-S)

ABSTRACT: **The objective of the present study was to elaborate a survival model that integrates anatomic factors, according to the 2010 seventh edition of the tumour, node and metastasis (TNM) staging system, with clinical and molecular factors.**

Pathologic TNM descriptors (group A), clinical variables (group B), laboratory parameters (group C) and molecular markers (tissue microarrays; group D) were collected from 512 early-stage nonsmall cell lung cancer (NSCLC) patients with complete resection. A multivariate analysis stepped supervised learning classification algorithm was used.

The prognostic performance by groups was: areas under the receiver operating characteristic curve (C-index): 0.67 (group A), 0.65 (Group B), 0.57 (group C) and 0.65 (group D). Considering all variables together selected for each of the four groups (integrated group) the C-index was 0.74 (95% CI 0.70–0.79), with statistically significant differences compared with each isolated group (from $p=0.006$ to $p<0.001$). Variables with the greatest prognostic discrimination were the presence of another ipsilobar nodule and tumour size >3 cm, followed by other anatomical and clinical factors, and molecular expressions of phosphorylated mammalian target of rapamycin (phospho-mTOR), Ki67cell proliferation index and phosphorylated acetyl-coenzyme A carboxylase.

This study on early-stage NSCLC shows the benefit from integrating pathological TNM, clinical and molecular factors into a composite prognostic model. The model of the integrated group classified patients with significantly higher accuracy compared to the TNM 2010 staging.

KEYWORDS: Ki67cell proliferation index, lung cancer, mammalian target of rapamycin, prognosis, staging

L ung cancer is the leading cause of death in Spain, accounting for 20,000 deaths in 2007 [1]. The best survival rates are in patients with early-stage nonsmall cell lung cancer (NSCLC) who undergo complete resection. However, only a small percentage of patients undergo surgical treatment and, even in the best-case scenario (stages pIA and pIB), in Spain, $>40\%$ of patients die within 5 yrs following resection [2].

In addition, the 2010 tumour, node and metastasis (TNM) classification has only been given a coefficient determination value (r^2) of <0.30 [3], thereby leaving most of the prognostic variance unexplained.

In the last 20 yrs, an increase in the publications on the prognosis of NSCLC has been detected [4]. Most of these publications focus on factors associated with the tumour, with special emphasis on prognostic molecular factors. The observation of several problems has prompted the appearance of recommendations for the study of prognostic factors in malignant tumours, including to conduct prognostic studies using immunohistochemistry [5].

Since 2006, several costly and complex prognostic classification systems for NSCLC have been gradually proposed, based on genetic or epigenetic

This article has supplementary material available from www.erj.ersjournals.com

molecular information, with miscellaneous study methodology. Despite this intense investigation, a scarce reproducibility of the different studies has been observed with regard to the selection of a few markers [6–8]. Among other problems, in most cases, variables of anatomic extent (TN descriptors) were deficiently treated in the models and did not specify the biases related to the selection of the study population.

The main objective of this study was to construct a composite prognostic survival model integrating the anatomic extent of the tumour with clinical, functional and molecular factors in a clearly defined population of patients with early-stage NSCLC.

METHODS
The study population included patients belonging to the Bronchogenic Carcinoma Cooperative Group of the Spanish Society of Pneumology and Thoracic Surgery (GCCB-S). There were 2,994 patients prospectively collected between 1993 and 1997. These patients are part of the international database used by the International Association for Study of Lung Cancer (IASLC) to update the TNM classification of lung cancer, the seventh edition of which was published in 2009 [9].

A total of 512 patients with NSCLC in pathologic (p) stages I–II, who underwent complete resection in six hospitals randomly selected among the 19 hospitals of the GCCB-S were included in this study. The seventh edition of the TNM classification was used for tumour staging [9].

Surgical specimens were studied following a standard protocol [10]. Histological types were independently established by three pathologists (F. López-Ríos, E. Conde, and A. Suárez-Gauthier (all Pathology Dept, Hospital Universitario 12 de Octubre, Madrid, Spain)) according to the World Health Organization 2004 classification [11]. All discrepancies were resolved by consensus.

A sample size of ∼500 patients was considered adequate for the expected presence of a 55–60% death rate in a 5-yr interval from time zero of calculation of survival, and about 25–35 variables on multivariate analysis.

The Institutional Review Board (Hospital Universitario 12 de Octubre) approved the protocols, and written consent was obtained from all the subjects of this study.

Initial available variables (>200) were included in four different groups: the TNM histology group (group A), which contained all qualitative and quantitative descriptors that define each TN category of stages pI–II; the clinical variables group (group B); the analytical and functional variables group (group C); and the molecular variables group (group D), which included 32 markers that explored five biochemical pathways (see online supplementary material).

Several steps were undertaken to build the predictive model. First, in each group, univariate analysis for selection of significant prognostic variables was performed by the Kaplan–Meier method. A p-value <0.3 was chosen as threshold for selection. Secondly, with the variables selected for each group, a classification tree was built by supervised learning classification algorithm. We consider vital status at 5-yr survival as the dependent variable at each terminal node of the classification tree. This was followed by multivariate

analysis by recursive partitioning decision tree using the supervised learning classification algorithm C4.5 constructed with R interface to Weka [12]. Each group had a tree with several terminal nodes. Every terminal node had a different probability of overall 5-yr survival. Group terminal nodes with minimal and maximal probababilities are shown. Thirdly, an integrated group was built with the variables obtained in the second step for all groups. Finally, a 5-yr probability of survival (Kaplan–Meier) was calculated (see Methods in the online supplementary material) for the clinical pattern of variables obtained in each terminal node of the integrated group.

The model's ability to discriminate between patients with or without the event was assessed using the area under the receiver operating characteristic (ROC) curve (AUC) method, measured by the concordance index (C-index) [13], and its overall predictive capacity, measure by the coefficient of determination [14]. Stata version 10 (StataCorp, College Station, TX, USA) was used for the remaining results. Given the digit preference in the tumour size variable, Schoenfeld's procedure was used [15]. (see online supplementary material). The internal validation of the model's estimation was calculated by bootstrapping.

RESULTS
Mean follow-up time for the cohort was 120 months. Median age was 67 yrs, with a mean $\pm$ SD age of 65.5 ± 8.3 yrs. The basic descriptive data of this series of 512 patients are shown in table 1. All data for the variables considered are stated in the online supplementary material.

TABLE 1	Basic descriptive data
Clinical data	**n (%)**
Male sex	474 (92.6)
Active smoker	274 (53.5)
Previous tumour	93 (18.2)
COPD	234 (45.7)
Arterial hypertension	90 (17.6)
Performance status ECOG grade 0–1	501 (97.8)
TNM staging	
pT1	107 (20.9)
pT2	365 (71.3)
pT3	40 (7.8)
pN0	430 (84)
pN1	82 (16)
Histological type	
Squamous cell carcinoma	324 (63.3)
Adenocarcinoma	117 (22.9)
Large cell carcinoma	62 (12.1)
Others	9 (1.8)
Treatment-related data	
Pneumonectomy	114 (22.3)
Lobectomy or bilobectomy	336 (65.6)
Sublobar resections or combination	62 (12.1)

For more information, see Table E-1 in the online supplementary material. COPD: chronic obstructive pulmonary disease; ECOG: Eastern Cooperative Oncology Group; TNM: tumour, node and metastasis; p: pathologic stage.

Table 2 shows, for each group of variables, the 30 variables selected after univariate analysis of the prognosis of survival, using the pre-established $p \leqslant 0.3$ statistical significance limit. Upon application of Schoenfeld's procedure, tumour size was

TABLE 2	Variables selected in the univariate analysis by group	
Variable	**Affected cases[#] n**	**Log rank p-value**
Group A		
Visceral pleura	113	0.0053
Parietal pleura	13	0.038
Tumour size		0.003
Proximal bronchus	150	0.05
Nodule in the same lobe	13	0.035
Atelectasis-pneumonitis	374	0.30
pN1	79	0.04
pTdi	34	0.013
Squamous cell carcinoma	324	0.21
Low tumour differentiation	32	0.044
Group B		
Previous tumour	92	0.03
Active smoker	277	0.085
Cardiac ischaemic disease	35	0.1
Arterial hypertension	89	0.06
COPD	235	0.14
Any comorbidity	295	0.006
Performance status ECOG grade 3–4	7	0.04
Upper age tercile	165	0.09
Group C		
Lower haemoglobin tercile	176	0.02
Lower FEV1 tercile	172	0.02
Lower FVC tercile	170	0.04
Group D		
Cell cycle		
P27	270	0.25
Ki67	353	0.30
Apoptosis		
Survivin-C	107	0.08
NF-κB	140	0.19
Adhesion molecules		
E-cadherin	140	0.04
β-catenin	12	0.24
Signal receptor–transducers		
Phospho-mTOR	261	0.27
Phospho-ACC	217	0.14
Others		
P63	277	0.15

p: pathologic stage; N: node; T: tumour; di: directly invades diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein; COPD: chronic obstructive pulmonary disease; ECOG: Eastern Cooperative Oncology Group; FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity; NF: nuclear factor; mTOR: mammalian target of rapamycin; ACC: acetyl-coenzyme A carboxylase. [#]: for molecular variables, the number of cases corresponds to positive cases for that marker.

distributed in three prognostic strata: 0–3, 3.1–7 and >7 cm. In bronchial involvement, the proximal location of the tumour distinguishes two groups: the most distal one, with endoscopic location at the level of the segmental bronchi or more distal bronchi; and the proximal one, with lobar or main bronchial location >2 cm from the tracheal carina.

Multivariate analysis selected different independent prognostic factors, with diverse interdependence, for each group of variables. Only 30 patients (5.8%) did not have adequate follow-up. Table 3 describes that selection, by group, showing the AUC associated for those variables. The probability spectrum of the event in each decision tree was different amongst these groups (table 3).

Multivariate analysis by classification tree of the entire set of variables selected from all the groups (integrated group) obtained five descriptor variables of the pTN group (group A; another nodule in the same lobe of the primary tumour and tumour size strata first, and involvement of other thoracic structures, level of endobronchial location, and presence of atelectasis or pneumonitis), four clinical variables (group B; performance status, active smoker, arterial hypertension and age) and three molecular variables (group D; phosphorylated mammalian target of rapamycin (phospho-mTOR), Ki67 and phosphorylated acetyl-coenzyme A carboxylase (phospho-ACC)). A significant improvement was identified (p<0.001 to p=0.006) in the integrated group AUC (all variables of all groups; AUC 0.74, 95% CI 0.70–0.79) over the previously described AUC values for that parameter, considering each group independently (table 4 and fig. 1). The probability spectrum of overall 5-yr survival also increased in the integrated group from 0.16 to 0.80 (a 64% difference; fig. 2). The coefficient of determination (r^2) was 0.24.

The internal validation of the final model was assessed by the bootstrap resampling technique. The average apparent AUC was 0.74, which was expected (based on bootstrapping) to decrease from 0.08 to 0.66. Figure 2 shows the interdependence and hierarchy over the discriminatory power of each variable of the integrated group. In the model, it was observed that the presence of another nodule in the same lobe of the primary tumour bears the maximum discriminatory capacity.

Given the patterns obtained in each node of the tree-based integrated group, overall 5-yr survival was calculated for each pattern. This allowed us to see the range of probability of survival according to the patients' clinical pattern. Given the prognostic similarity of some branches of the tree-based model, some of them have been combined according to their probability of 5-yr survival into four groups: group 1 (n=165), with probability of 5-yr survival of 0.75 (95% CI 0.68–0.81); group 2 (n=92), with a probability of 5-yr survival of 0.64 (95% CI 0.54–0.73); group 3 (n=83), with a probability of 5-yr survival of 0.40 (95% CI 0.29–0.50); and, finally, group 4 (n=142), with a probability of 5-yr survival of 0.25 (95% CI 0.18–0.32) (fig. 3).

DISCUSSION
Summary of main data
This prognostic, multivariate, multidimensional and multicentre analysis of 482 patients with completely resected early-stage (pI–II) NSCLC selected the presence of another nodule in

TABLE 3	Selection of variables in the multivariate analysis using supervised learning classification method and spectrum of probabilities of overall 5-yr survival (extreme values)		
Group	**Selected variables**	**AUC C index (95% CI)**	**Overall 5-yr survival extreme values**
A	Nodule in the same lobe, tumour size, pTdi, proximal bronchus and atelectasis–pneumonitis	0.67 (0.62–0.71)	0.33–0.86
B	Arterial hypertension, age, performance status, active smoker, previous tumour and COPD	0.65 (0.60–0.70)	0.26–0.77
C	Haemoglobin	0.57 (0.54–0.60)	0.44–0.70
D	Phospho-ACC, Ki67, P63, E-cadherin, phospho-mTOR, P27 and NF-κB	0.65 (0.60–0.70)	0.25–0.72

AUC: area under the curve; C-index: concordance index; p: pathologic stage; T: tumour; di: directly invades diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein; COPD: chronic obstructive pulmonary disease; ACC; acetyl-coenzyme A carboxylase; mTOR: mammalian target of rapamycin; NF: nuclear factor.

the same lobe of the primary tumour and the tumour size as the most discriminative factors with regard to survival. Other selected factors included clinical variables (performance status, active smoking, presence of arterial hypertension and age), anatomical variables (involvement of thoracic structures, presence of atelectasis or pneumonitis and level of endobronchial location), analytic variables (haemogoblin) and some molecular markers (phospho-mTOR, Ki67, and phospho-ACC). The integration of tumour extent, clinical and molecular factors (integrated group) significantly improved the discriminatory ability of the model compared with its ability to discriminate when these groups of factors were analysed individually.

This integration of factors reached an AUC of 0.74 (95% CI 0.70–0.79) and obtained an r^2 coefficient of 0.24; both data indicate the need for further research to improve prognostic capacity for NSCLC in its early stages. The most extreme limits of the prognostic spectrum observed showed the probability of survival at 5 yrs to be 0.16–0.80: a 64% difference. This difference was greater than that described in 2009 by the new IASLC/International Union Against Cancer/American Joint Committee on Cancer lung cancer staging classification [16] for patients with stage pIA and pIIB tumours: a 37%

TABLE 4	Comparison of receiver operating characteristic area under the curve (AUC) for each group of variables in relation to the integrated group AUC, taking all groups into account			
Group	**AUC C-index**	**SE**	**p-value[#]**	**r^2**
A	0.6673	0.024	0.0002	0.1250
B	0.6524	0.024	0.0058	0.1007
C	0.5717	0.017	<0.001	0.0493
D	0.6497	0.025	0.0035	0.1039
Integrated group	0.7438	0.022		0.2382

n=482. C-index: concordance index. [#]: comparison of the integrated group AUC (taking all variables from all groups into account) with the AUC of group A, B, C or D.

difference. The 2010 TNM classification has only been given a coefficient determination value (r^2) of <0.30, despite the great certainty it offers in the classificaiton of the prognosis of death, as shown by the 40% of patients with stage IV tumours at the time of diagnosis [3].

AUC in the integrated group

In the last 10–15 yrs, most publications of a genetic nature, clinical–genomic mixed models, and calculations with epigenetic or proteomic studies have shown that the combination of anatomical extent variables with molecular biology variables improves prognostic discrimination in an independent fashion [7, 17–19].

With different outcomes, and several types of NSCLC populations and study platforms, diverse publications have reported AUC 0.58–0.75 on most occasions [7, 18, 20, 21], even though in some population subsets, these AUC are higher [17, 20]. On other occasions, the image of the ROC curve graphically depicts results excellently, even though the
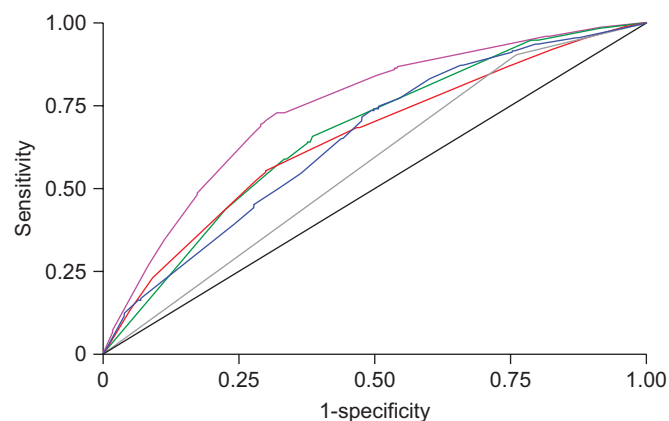

FIGURE 1. Area under the receiver operating characteristic curves (AUC) for each group of variables, taking all groups into account. Black: reference; green: group A (anatomical and histological variables; AUC 0.67); red: group B (clinical variables; AUC 0.65); grey: group C (functional and laboratory variables; AUC 0.57); blue: group D (molecular variables from tissue microarrays; AUC 0.65); purple: integrated group (groups A, B, C and D combined; AUC 0.74).
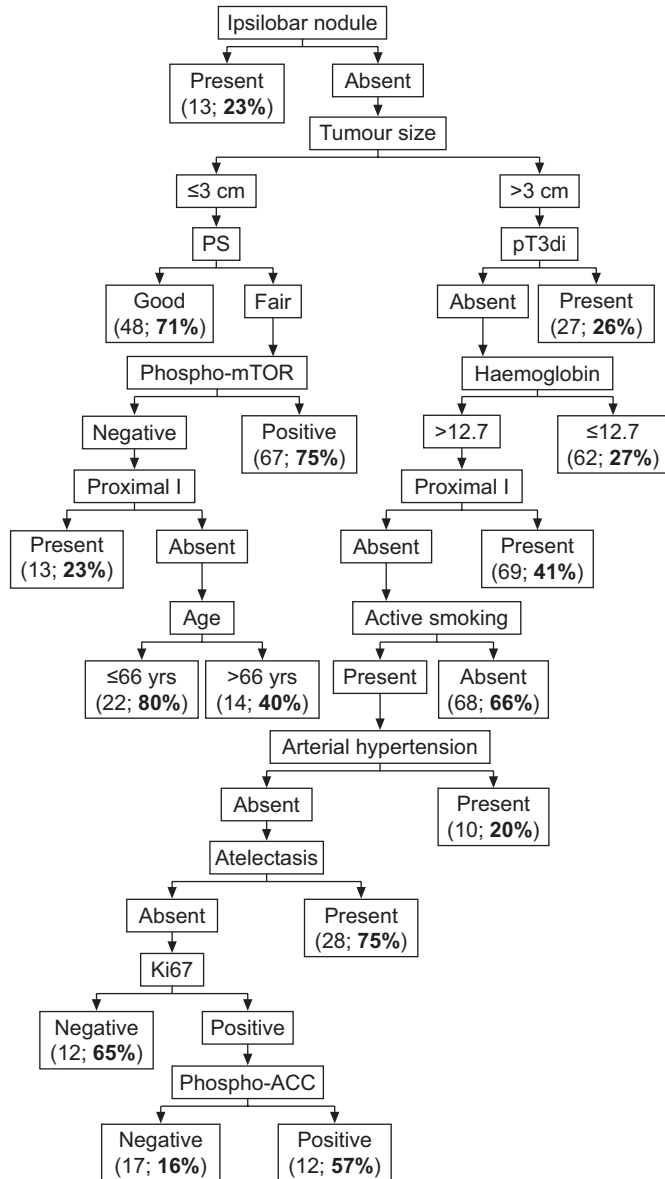
FIGURE 2. Classification tree (total n=482). Data are presented as the number of cases and their probability of overall 5-yr survival (in bold text) for each terminal node. PS: performance status; phospho-mTOR: phosphorylated mammalian target of rapamycin; p: pathologic stage; T3di: directly invades diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein; proximal I: proximal bronchial involvement; atelectasis: atelectasis–pneumonitis; ACC: acetyl-coenzyme A carboxylase.

FIGURE 3. Survival curves of overall survival (5 yrs) for terminal node groups of similar survival from the classification tree. Black: group 1 (n=165; terminal node case grouping with overall 5-yr survival 71–81%); red: group 2 (n=92; terminal node case grouping with overall 5-yr survival 57–66%); green: group 3 (n=83; terminal node case grouping with overall 5-yr survival 40–41%); blue: group 4 (n=142; terminal node case grouping with overall survival 16–27%).

findings suggested ''that the clinical covariates should be collected with the same care as used for obtaining gene expression signatures'' [19]. In that study, with overall integration of all variables, the C-index (AUC) varied by hospital and type of classifier (study method) from 0.61 to 0.76 (for all stages), and from 0.51 to 0.80 for stage I, with a maximum prognostic spectrum of survival at 5-yrs (extremes) of 50% (considering all stages, using an overall integrated method, in one single centre, and using gene cluster and ridge regression analysis) [19].

Prognostic spectrum

The prognostic spectrum reached in our study with the overall integrated model presents a 64% difference between the 5-yr survival extremes in a population of patients with completely resected stage I–II NSCLC. This spectrum is similar or superior to that reached in other studies, which employed much more complex and costly molecular studies [6–8, 18–21], and clearly inferior to the 75–80% values of other studies [17, 22, 23].

TNM descriptors and clinical variables

In our final model (fig. 2), the presence of another nodule in the same lobe of the primary tumour presented a 5-yr survival of 23%, similar to the 5-yr survival reported in the seventh edition of the TNM classification for T descriptors (another nodule in the same lobe; any R; any pN) [24]. The same happened with high values of tumour size, taking into account T descriptors alone [24].

Performance status is a recognised prognostic factor in lung cancer. Being an active smoker at the time of diagnosis and treatment of NSCLC is an independent prognostic factor *versus* not having been a smoker or being an ex-smoker, with such an effect not being necessarily explained by associated tobacco-related comorbidity [25]. To our knowledge, there is no published information about arterial hypertension as a prognostic factor in lung cancer. Finally, within the group of clinical variables, age has already been established as an independent prognostic factor when gene signatures are taken into account [23].

quantification of its area is not shown [22]. Our C-index value or AUC of 0.74 (95% CI 0.70–0.79) is within the range of reported values.

In a study from the Consortium for the Molecular Classification of Lung Adenocarcinoma, a total of 442 cases of lung adenocarcinomas was analysed, and gene expression was integrated with other pathological and clinical data [19]. Using any method of analysis or study of NSCLC, and in different institutions, the addition of clinical covariates improved the hazard ratio of gene expression to a point where it became statistically significant. The authors concluded that their

Molecular variables

The first molecular component selected in this study was phospho-mTOR (fig. 2). Within the different molecular pathways of NSCLC, the phosphoinositide 3-kinase/AKT pathway has received a lot of attention because of its involvement in cell proliferation, and in invasion and apoptosis mechanisms [26]. This pathway is frequently over-activated in NSCLC. Phospho-mTOR is directly involved in tumour proliferation. Phospho-mTOR activation is of clinical interest, given the possibility of using specifically targeted therapies.

The Ki67 cell proliferation index was selected at a later stage in the modelling process (fig. 2). Ki67 is a DNA-binding nuclear protein that is present in all phases of the cell cycle, except in the quiescent G0 phase, which can be easily indentified by immunohistochemistry. Its expression is associated with prognosis of the cancer patient and, specifically, of those with NSCLC. A recent systematic review and meta-analysis concluded that Ki67 was associated with bad prognosis in NSCLC, although, in stages I–II, with over 1,000 patients from eight different studies, no statistically significant hazard ratio was found [27].

Finally, in the subgroup of patients with expression of Ki67, expression of phospho-ACC had little prognostic value. This observation, which has been scarcely studied, had been previously detected [28].

Limitations and strengths

This study is both negative and positive. It is negative because the discriminative capacity of this model (C-index 0.74) implies that there is room for improvement, and it is positive because it demonstrates that all variables (anatomic tumour extent, clinical, molecular, *etc.*) are important, and that there is a clinically relevant use for each and every one of them.

This study presents several limitations. One of the selected outcomes (overall survival) includes death from any cause, which can result in underestimating the biological–molecular prognostic factors associated with NSCLC. However, in an integrated, multidimensional, prognostic approach, clinical factors, as evidenced by this study, may be selected as prognostic factors if all causes of death are considered.

The limitations of the molecular analyses in the present study are derived from the procedure used: tissue microarrays and immunohistochemical study [5]. The online supplementary material provides a detailed description of the procedures used and of the controls performed, including an interobserver analysis and intercore agreement.

The strengths of this work lie in the size of the studied population (n=482), its definition and selection, and in the quality controls performed for all types of variables, including anatomic (each internal descriptor of pT and pN), clinical and molecular variables. It consists of a series of consecutive cases with prospective collection of all variables in several centres that share the same tumour and therapeutic classification: NSCLC, stages with maximum certainty (pathologic staging) and early stages (stages pI–II), with adequate pathological mediastinal lymph node staging and complete resection. It is therefore a homogeneous population, which would, in theory,

facilitate its potential reproducibility in other areas and corrects the so-called ''denominator effect in survival'' [2].

Multivariable analysis using classification and decision tree

For the objectives of our study, it is helpful to consider all types of variables, regardless of the number of times that these variables have been studied in all cases, and to understand the hierarchy and relationship between the different prognostic factors selected. It therefore consists of a very intuitive explanatory model that explores interactions and conditioning between factors.

The results measured by the C-index are modest, but similar to the results obtained in other recent similar studies [19]. They also are less expensive than gene expression-based prognostic signatures for NSCLC, which have not yet proved better clinical utility [29].

REFERENCES

1 Instituto Nacional de Estadística. Death according to Cause of Death 2007. www.ine.es/en/inebmenu/mnu_salud_en.htm#3 Date last accessed: February 1, 2010.

2 Duque J, López-Encuentra A, Rami-Porta R, *et al.* Survival of 2,991 patients with surgical lung cancer: the denominator effect in survival. *Chest* 2005; 28: 2274–2281.

3 Goldstraw P, Crowley J, Chansky K, *et al.* The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edn of the TNM Classification of malignant tumours. *J Thorac Oncol* 2007; 2: 706–714.

4 Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest* 2002; 122: 1037–1057.

5 Zhu CQ, Shih W, Ling CH, *et al.* Immunohistochemical markers of prognosis in non-small cell lung cancer: a review and proposal for a multiphase approach to marker evaluation. *J Clin Pathol* 2006; 59: 790–800.

6 Guo NL, Wan YW, Tosun K, *et al.* Confirmation of gene expression-based prediction of survival in non-small cell lung cancer. *Clin Cancer Res* 2008; 14: 8213–8220.

7 Lau SK, Boutros PC, Pintilie M, *et al.* Three-gene prognostic classifier for early-stage non small-cell lung cancer. *J Clin Oncol* 2007; 25: 5562–5569.

8 Roepman P, Jassem J, Smit EF, *et al.* An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clin Cancer Res* 2009; 15: 284–290.

9 Goldstraw P, ed. International Association for the Study of Lung Cancer. Staging Handbook in Thoracic Oncology. Orange Park, Editorial Rx-Press, 2009.

10 Lester SC. Manual of Surgical Pathology. Philadelphia, Churchill Livingstone, 2001.

11 Travis WD, Brambilla E, Müller-Hermelink HK, *et al.* Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. World Health Organization Classification of Tumours. Lyon, IARC Press, 2004.

12 Quinlan R. C4.5: Programs for Machine Learning. San Mateo, Morgan Kaufmann Publishers, 1993.

13 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.

14 Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika* 1991; 78: 691–692.

15 Schoenfeld DA. Analysis of categorical data: logistic model. *In*: Mike V, Stanley KE, eds. Statistics in medical research. New York, Wiley, 1982; pp. 443–454.

16 Detterbeck FC, Boffa DJ, Tanoue LT. The new lung cancer staging system. *Chest* 2009; 136: 260–271.

17 Potti A, Mukherjee S, Petersen R, *et al.* A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 2006; 355: 570–580.

18 Lee ES, Son DS, Kim SH, *et al.* Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin Cancer Res* 2008; 14: 7397–7404.

19 Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JM, *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; 14: 822–827.

20 Guo L, Ma Y, Ward R, *et al.* Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clin Cancer Res* 2006; 12: 3344–3354.

21 Sun Z, Wigle DA, Yang P. Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival. *J Clin Oncol* 2008; 26: 877–883.

22 Zhu ZH, Sun BY, Ma Y, *et al.* Three immunomarker support vector machines-based prognostic classifiers for stage IB non-small-cell lung cancer. *J Clin Oncol* 2009; 27: 1091–1099.

23 Chen HY, Yu SL, Chen CH, *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; 356: 11–20.

24 Rami-Porta R, Ball D, Crowley J, *et al.* The IASLC Lung Cancer Staging Project: proposals for the revision of the T descriptors in the forthcoming (seventh) edn of the TNM classification for lung cancer. *J Thorac Oncol* 2007; 2: 593–602.

25 Tammemagi CM, Neslund-Dudas C, Simoff M, *et al.* Smoking and lung cancer survival: the role of comorbidity and treatment. *Chest* 2004; 125: 27–37.

26 Solomon B, Pearson RB. Class IA phosphatidylinositol 3-kinase signalling in non-small cell lung cancer. *J Thorac Oncol* 2009; 4: 787–791.

27 Martin B, Paesmans M, Mascaux C, *et al.* Ki-67 expression and patients survival in lung cancer: systematic review of the literature with meta-analysis. *Br J Cancer* 2004; 91: 2018–2025.

28 Conde E, Suarez-Gauthier A, García-García E, *et al.* Specific pattern of LKB1 and phospho-acetyl-CoA carboxylase protein immunostaining in human normal tissues and lung carcinomas. *Hum Pathol* 2007; 38: 1351–1360.

29 Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst* 2010; 102: 1–11.