ERJ Express. Published on September 3, 2010 as doi: 10.1183/09031936.00028610

COMPOSITE ANATOMIC-CLINICAL-MOLECULAR PROGNOSTIC MODEL IN NON-SMALL-CELL LUNG CANCER

A multicenter study of 512 patients

Angel López-Encuentra[1], Fernando López-Ríos[2], Esther Conde[2], Ricardo García-Luján[1], Ana Suárez-Gauthier[2], Nuria Mañes[3], Guadalupe Renedo[4], José Luis Duque-Medina[5], Elena García-Lagarto[6], Ramón Rami-Porta[7], Guadalupe González-Pont[8], Julio Astudillo-Pombo[9], José Luis Maté-Sanz[10], Jorge Freixinet[11], Teresa Romero-Saavedra[12], Monserrat Sánchez-Céspedes[13], Agustin Goméz de la Camara[14], on behalf of the Bronchogenic Carcinoma Cooperative Group of the Spanish Society of Pneumology and Thoracic Surgery (GCCB-S) [15]

[1] Pneumology Department, Hospital Universitario 12 de Octubre, and Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERes), Madrid, Spain; [2] Pathology Department, Hospital Universitario 12 de Octubre, Madrid, Spain; [3] Thoracic Surgery Department, Fundación Jiménez Díaz, Madrid, Spain; [4] Pathology Department, Fundación Jiménez Díaz, Madrid, Spain; [5] Thoracic Surgery Department, Hospital Clínico Universitario, Valladolid, Spain; [6] Pathology Department, Hospital Clínico Universitario, Valladolid, Spain; [7] Thoracic Surgery Service, Hospital Universitario Mutua de Terrassa, Terrassa, Barcelona, Spain; [8] Pathology Department, Hospital Universitario Mutua de Terrassa, Terrassa, Barcelona, Spain; [9] Thoracic Surgery Department, Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain; [10] Pathology Department, Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain; [11] Thoracic Surgery Department, Hospital Dr. Negrin, Las Palmas, Spain; [12] Pathology Department, Hospital Dr. Negrin, Las Palmas, Spain; [13] Molecular Pathology Department, Spanish National Cancer Centre, Madrid, Spain; [14] Clinical Epidemiology Unit, Hospital Universitario 12 de Octubre, CIBERes, Madrid, Spain; [15] Coordinators and members of the GCCB-S: See Appendix.

Corresponding author: Angel López-Encuentra, MD, PhD. Pneumology Service, Hospital Universitario 12 de Octubre, Avenida de Córdoba s/n, 28041 Madrid, Spain. FAX: + 34 91 3908492. E-mail: lencuent@h12o.es

Keywords: Ki67cell proliferation index, lung cancer, mammalian target of rapamycin (mTORp), prognosis, staging**.**

Word count body text: 2,729

Word count abstract: 199

This article has an online supplement, which is accessible from this issue´s table of contents at **ERJ-online** or our website www.mbeneumologia.org

ABSTRACT

The objective is to elaborate a survival model that integrates anatomic factors, according to the 2010 seventh edition of the tumour-node-metastasis (TNM) staging, 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 NSCLC with complete resection. A multivariate analysis steped supervised learning classification algorithm was used.

The prognostic performance by groups is: areas under the ROC curve (C-index): 0.67 (Group A), 0.65 (Group B), 0.57 (Group C) and 0.65 (Group D). Considering together all variables selected for each of the 4 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 are the presence of another ipsilobar nodule and tumour size >3 cm; followed by other anatomic and clinical factors; and molecular expressions of mammalian target of rapamycin (phospho-mTOR), Ki67cell proliferation index and p-Acetil-CoA-Carboxylase.

This study on early NSCLC shows the benefit from integrating pTNM, 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.

INTRODUCTION

Lung cancer is the leading cause of death in Spain, accounting for 20,000 deaths in 2007.[1] The best survival occurs in patients with early stage non-small 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), more than 40% of patients die within 5 years following resection in Spain.[2]

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

In the last 20 years an increase in the publications on the prognosis of NSCLC has been detected.[4] Most 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 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.

On a clearly defined population of patients with early stage NSCLC, 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.

METHODS *(complete methods are provided in an online supplement)*

The study population included patients pertaining 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 (FL-R,EC,AS-G) according to the World Health Organisation 2004 classification.[11] All discrepancies were resolved by consensus.

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

Initial available variables (more than 200) are included in 4 different groups: The TNM-histology Group contains all and every single qualitative and quantitative descriptors that define each T-N category of stages pI-II, the Group of clinical variables, the Group of analytical and functional variables and the Group of molecular variables that includes 32 markers that explore five biochemical pathways (online supplement).

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. Second, with the variables selected for each Group, a classification tree was built by supervised learning classification algorithm. We consider vital status at 5-year survival as 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 has a different probability of overall 5-year survival. Group terminal nodes with minimal and maximal probabababilities are displayed. Third, an Integrated Group was built with the variables obtained in the second step for all Groups. Fourth a five-year probability of survival (Kaplan Meier) was calculated (see Methods in online supplement) for the clinical pattern of variables obtained in each terminal node of the Integrated Group.

The model's ability to discriminate amongst patients with or without the event was assessed using the area under the receiver operative characteristics (ROC) curve (AUC) method, measured by the concordance index (C index).[13] and its overall predictive capacity with the coefficient of determination.[14] The STATA programme was used for the remaining results.[15] Given the digit preference in the tumour size variable, Schoenfeld's procedure was used.[16] (see online supplement). The internal validation of the model's estimation is calculated by bootstrapping.

RESULTS

Mean follow-up for this cohort was 120 months. Median age was 67 years, with a mean age of 65.5 (SD±8.3). The basic descriptive data of this series of 512 patients are shown on Table 1. All data for all considered variables are stated in the online supplement.

Table 2 shows, for each group of variables, the 30 variables selected after univariate analysis on the prognosis of survival with the $p \leq 0.3$ pre-established statistical signification limit. Upon application of Schoenfeld's procedure, tumour size was distributed in three prognostic strata: 0-3 cm, 3.1-7 cm, >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 at more than 2 cm from the tracheal carina.

Multivariate analysis selected different independent prognostic factors with diverse interdependence amongst them 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 value under the ROC curve (AUC) associated to such variables. The probability spectrum of the event in each decision tree was different amongst these groups (Table 4).

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 (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 (performance status, active smoker, arterial hypertension, age) and three molecular variables (phospho-mTOR [mTORp], Ki67 and p-Acetil-CoA-Carboxylase [p-ACC]). A significant improvement is identified ($p<0.001$ – $p=0.006$) in the Integrated Group AUC (all variables of all groups) (0.74 [95%CI: 0.70-0.79]) over the previously described AUC values for that parameter considering each group independently (Table 5) (Figure 1). The probability spectrum of overall 5-year survival also increased in the Integrated Group from 0.16 to 0.80 (a 64% difference) (Figure 2). The coefficient of determination (R^2) was 0.24.

The internal validation of the final model was assessed by the bootstrap re-sampling technique. The average apparent ROC area was 0.74, which was expected (based on bootstrapping) to decrease 0.08 down to 0.66. Figure 2 shows the interdependence and hierarchy over the discriminatory power of each variable of the Integrated Group. In this model it is 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-years survival (OS) 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-year survival into four groups: Group 1 (n=165), with a probability of 5 years OS of 0.75 [95%CI: 0.68 – 0.81]; group 2 (n=92), with a probability of 5 years OS of 0.64 [95%CI: 0.54 – 0.73]; group 3 (n=83), with a probability of 5 years OS of 0.40 [95%CI: 0.29 – 0.50]; and lastly, group 4 (n=142), with a probability of 5 years OS of 0.25 [95%CI: 0.18 – 0.32] (Figure 3).

DISCUSSION

Summary of main data

This prognostic, multivariate, multidimensional and multicenter analysis on 482 patients with completely resected early stage (pI-II) NSCLC selected the presence of another nodule in the same lobe of the primary tumour and the tumour size as the most discriminative factor with regard to survival.Other selected factors include clinical variables (performance status, active smoking, presence of arterial hypertension, and age), other variables of anatomic extent (involvement of thoracic structures, presence of atelectasis or pneumonitis, level of endobronchial location), analytic variables (haemogoblin), and some molecular expressions (phosphor-mTORp, Ki67, and p-ACC). The integration of tumour extent, clinical and molecular factors (Integrated Group) significantly improves the discriminative ability of the model compared with the abitily to discriminate when these groups of factors are analyzed individually.

This integration of factors reaches an area under the curve of 0.74 (95%CI:0.70-0.79) and obtains a 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 show the probability of survival at 5 years to be between 0.16 and 0.80: a 64% difference. This difference is greater than that described in 2009 with the new IASLC-

International Union Against Cancer (UICC) - American Joint Committee on Cancer (AJCC) lung cancer staging classification [17] for patients with stage pIA and pIIB tumours: a 37% difference. The 2010 TNM classification has only been given a coefficient determination value (R^2) of less than 0.30, despite the great classificatory certainty in the prognosis of death shown by the 40% of patients with stage IV tumours at the time of diagnosis.[3]

Area under the curve ROC in Integrated Group.

In the last 10 to 15 years, most publications of a genetic nature, clinical-genomic mixed models, calculations with epigenetic or proteomic studies have shown that the combination of anatomic extent variables with molecular biology variables improves prognostic discrimination in an independent fashion.[7,18-20]

With different outcomes, several types of NSCLC populations and study platforms, diverse publications have reported areas under the curve (AUC) between 0.58 and 0.75 on most occasions [7,19,21,22], even though in some population subsets, these AUC are higher.[18,21] On other occasions the image of the ROC curve depicts excellent results graphically even though the quantification of its area is not shown.[23] Our concordance index value or AUC 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.[20] 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 findings suggested “that the clinical covariates should be collected with the same care as used for obtaining gene expression signatures”.[20] In the above mentioned experience with overall integration of all variables, the C index (AUC) varied per 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 years (extremes) of 50% (considering all stages, using an overall integrated method, in one single centre, and gene cluster and ridge regression analysis).[20]

Prognostic spectrum

The prognostic spectrum reached in our study with the overall integrated model presents a 64% difference between the 5-year 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 experiences which employed much more complex and costly molecular studies [6-8,19-22], and clearly inferior to the 75-80% values of other studies.[18,23,24]

TNM descriptors and clinical variables

In our final model (Figure 2) the presence of another nodule in the same lobe of the primary tumour had a possibility to present 5-year survival of 23%, similar to the 5-year survival reported in the seventh edition of the TNM classification for T descriptors (another nodule in the same lobe; any R; any pN).[25] The same happened with the high value of tumour size taking into account T descriptors alone.[25]

Performance status is a recognized 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.[26] 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.[24]

Molecular variables

The first molecular component selected in this study is the phosphorylated mammalian target of rapamycin (phospho-mTORp)(Figure 2). Within the different molecular pathways of NSCLC, the PI3K/AKT pathway has received a lot of attention because of its involvement in cell proliferation, and in invasion and apoptosis mechanisms.[27] This pathway is frequently over activated in NSCLC. phospho-mTORp is, within this pathway, directly related with tumour proliferation. phospho-mTORp activation has clinical interest given the possibility of using specifically targeted therapies.

The Ki67 cell proliferation index is selected at a later stage in the process (Figure 2). Ki67 is a DNA-binding nuclear protein that is present in all phases of the cell cycle of proliferating cells, except in the quiescent GO phase that can be easily studied by immunohistochemistry. Its expression is associated with prognosis of the cancer patient and, specifically, of those with NSCLC. A recent systematic review with meta-analysis concluded that Ki67 was associated to a bad prognosis in NSCLC, although, in stages I-II, with over 1,000 patients from 8 different studies, no statistically significant hazard ratio was found.[28]

Finally, in the subgroup of patients with positive expression of Ki67, expression of p-ACC has little prognostic value. This last observation, that has been scarcely studied, had been previously detected.[29]

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 much to be improved; 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 has been evidenced in this study, may be selected as prognostic factors if all causes of death are considered.

The limitations of the molecular study in our work are derived from the procedure used: tissue microarrays and immunohistochemical study.[5] The appendix (online supplement) 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 tumoral and therapeutic classification: NSCLC, stages with maximum certainty (pathologic staging) and early stages (pI-II stages) 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 experiences.[20] They also are less expensive than gene expression–based prognostic signatures for NSCLC, that have not proved, yet, a better clinical utility.[30]

Table 1

Basic descriptive data

(more information in Table E-1. Supplementary appendix)

Clinical data	**Frequency (%)**
Male sex	474 (92.6%)
Active smoker	274 (53.5%)
Previous tumour	93 (18.2%)
Chronic obstructive pulmonary disease	234 (45.7%)
Arterial hypertension	90 (17.6%)
Performance status (ECOG): Grade 0 or 1	501 (97.8%)
Staging pT - pN	
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%)

Table 2

Variables selected in the univariate analysis by Groups

GROUPS OF VARIABLES	**VARIABLE** (number of affected cases)(*)	**Log rank** (p)
Group A	Visceral pleura (113)	0.0053
pTN-descriptor and	Parietal pleura (13)	0.038
histological variables	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
Clinical variables	Active smoker (277)	0.085
	Cardiac ischemic disease (35)	0.1
	Arterial hypertension (89)	0.06
	Chronic obstructive pulmonary disease (COPD)(235)	0.14
	Comorbidity (any) (295)	0.006
	Performance status (ECOG 3-4) (7)	0.04
	Age (upper tercile) (165)	0.09

Table 2 (cont)

Variables selected in the univariate analysis by Groups

GROUPS OF VARIABLES	**VARIABLE** (number of affected cases)(*)	**Log rank** (p)
Group C	Haemoglobin (lower tercile)(176)	0.02
Analytical and functional	FEV1 (lower tercile) (172)	0.02
variables	FVC (lower tercile) (170)	0.04
Group D	Cell cycle	
Molecular variables	- P27 (270)	0.25
	- Ki67 (353)	0.30
	Apoptosis	
	- Survivin-C (107)	0.08
	- NFKβ (140)	0.19
	Adhesion molecules	
	- E-cadherin (140)	0.04
	- β-catenin (12)	0.24
	Signal receptors – transductors	
	- phospho-mTOR (261)	0.27
	- phospho-ACC (217)	0.14
	Others	
	- P63 (277)	0.15

(*) In molecular variables the number of cases reflected in the table correspond to positive cases for that marker

(**) pTdi: directly invades any of the following: diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein involvement.

Table 3

Selection of variables in the multivariate analysis by Groups using supervised learning classification method

GROUPS OF VARIABLES	**SELECTED VARIABLES**	**AUC; C index (*)** (95% CI)
Group A	Nodule in the same lobe	0.67
pTN-descriptor and	Tumour size	(0.62-0.71)
histological variables	pTdi (**)	
	Proximal bronchus	
	Atelectasis-pneumonitis	
Group B	Arterial hypertension	0.65
Clinical variables	Age	(0.60-0.70)
	Performance status	
	Active smoker	
	Previous tumour	
	COPD (***)	
Group C	Haemoglobin	0.57
Analytical and functional		(0.54-0.60)
variables		

Table 3 (cont)

Selection of variables in the multivariate analysis by Groups using supervised learning classification method

GROUPS OF VARIABLES	**SELECTED VARIABLES**	**AUC; C index (*)** (95% CI)
Group D	phospho-ACC	0.65
Molecular variables	Ki67	(0.60-0.70)
	P63	
	E-cadherin	
	phospho-mTOR	
	P27	
	NFKβ	

(*) Receiver Operative Characteristic (ROC). Area under the ROC curve (AUC); concordance index or C index
(**) pTdi: directly invades any of the following: diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein involvement.
(***) COPD: Chronic Obstructive Pulmonary Disease

Table 4

Spectrum of probabilities of overall five years survival (extreme values) by Groups

GROUPS OF VARIABLES	**SELECTED VARIABLES**	**OVERALL FIVE YEARS SURVIVAL (extreme values)**
Group A	Nodule in the same lobe	0.33 – 0.86
pTN-descriptor and	Tumour size	
histological variables	pTdi (*)	
	Proximal bronchus	
	Atelectasis-pneumonitis	
Group B	Arterial hypertension	0.26 – 0.77
Clinical variables	Age	
	Performance status	
	Active smoker	
	Previous tumour	
	COPD (***)	
Group C	Haemoglobin	0.44 – 0.70
Analytical and functional		
variables		

Table 4 (cont)

Spectrum of probabilities of overall five years survival (extreme values) by Groups

GROUPS OF VARIABLES	SELECTED VARIABLES	OVERALL FIVE YEARS SURVIVAL (extreme values)
Group D	phospho-ACC	0.25 – 0.72
Molecular variables	Ki67	
	P63	
	E-cadherin	
	phospho-mTOR	
	P27	
	NFKβ	

(**) pTdi: directly invades any of the following: diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein involvement.

Table 5

Comparison of ROC Areas for each Group of variables in relation to the Integrated Group ROC Area, taking all Groups into account

GROUP	**AUC; C index (*)**	**STANDARD ERROR**	**NUMBER OF CASES**	**p(**)**	**R^2**
Group A	0.6673	0.024	482	0.0002	0.1250
Group B	0.6524	0.024	482	0.0058	0.1007
Group C	0.5717	0.017	482	<0.001	0.0493
Group D	0.6497	0.025	482	0.0035	0.1039
Integrated Group	**0.7438**	**0.022**	482		**0.2382**

(*) Receiver Operative Characteristic (ROC): Area under the ROC curve; concordance index or C index

(**) Statistical signification: it is the comparison of the **Integrated Group** AUC (taking all variables from all Groups into account) with the AUC of each A, B, C or D group.

FUNDING: Supported by FIS grant 97/11, FEPAR-2000 grant, CIBER Respiratory Disease - ISCIII-CB06/06*,* CIBER Epidemiology and Public Health, FIS grant 03/46. This work has also been partially funded by the FMM, and has received financial aid from the Castilla-León regional government. The CIBER Enfermedades Respiratorias (www.ciberes.org) is an initiative of the Instituto de Salud Carlos III (ISCIII) of Spain.

ROLE OF FUNDING SOURCE: None of the institutions that have contributed financially to this paper have participated in its conception, design, analysis, interpretation of data, writing of its contents or in the decision to publish it.

CONFLICT OF INTEREST STATEMENT: The corresponding author (*Angel López-Encuentra*) confirms that he has full access to all the data of this study, and that he has the final responsibility for the decision to submit this manuscript for publication. All authors declare no conflict of interest with people or organisations that could inappropriately influence (bias) this work.

INSTITUCIONAL REVIEW BOARD: The Institutional Review Board approved the protocols, and written consent was obtained from all the subjects of this study.

REFERENCES

1- Death according to Cause of Death 2007. Instituto Nacional de Estadística. http://www.ine.es/en/inebmenu/mnu_salud_en.htm#3 (accessed 1 Feb 2010).

2- Duque J, López-Encuentra A, Rami-Porta R; Bronchogenic Carcinoma Cooperative Group of the Spanish Society of Pneumology and Thoracic Surgery. 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, Giroux DJ, Groome PA, Rami-Porta R, Postmus PE, Rusch V, Sobin L; International Association for the Study of Lung Cancer International Staging Committee; Participating Institutions. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition 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, Tsao MS. 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, Lin H, Msiska Z, Flynn DC, Remick SC, Vallyathan V, Dowlati A, Shi X, Castranova V, Beer DG, Qian Y. 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, Blackhall FH, Zhu CQ, Strumpf D, Johnston MR, Darling G, Keshavjee S, Waddell TK, Liu N, Lau D, Penn LZ, Shepherd FA, Jurisica I, Der SD, Tsao MS. Three-gene prognostic classificier for early-stage non small-cell lung cancer. *J Clin Oncol* 2007; 25: 5562-5569.

8- Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, Witteveen AT, Rzyman W, Floore A, Burgers S, Giaccone G, Meister M, Dienemann H, Skrzypski M, Kozlowski M, Mooi WJ, van Zandwijk N. 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 (Editor). International Association for the Study of Lung Cancer. Staging Handbook in Thoracic Oncology. Orange Park, Florida: Editorial Rx-Press 2009.

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

11- Travis WD, Brambilla E, Müller-Hermelink HK, Harris CC. 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, California: 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- StataCorp. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP 2007.

16- Schoenfeld DA. Analysis of categorical data: logistic model. In: Mike V, Stanley KE, editors. Statistics in medical research. New York, NY: Wiley 1982: 443-454.

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

18- Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, West M, Harpole DH Jr, Nevins JR. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 2006; 355: 570-580.

19- Lee ES, Son DS, Kim SH, Lee J, Jo J, Han J, Kim H, Lee HJ, Choi HY, Jung Y, Park M, Lim YS, Kim K, Shim Y, Kim BC, Lee K, Huh N, Ko C, Park K, Lee JW, Choi YS, Kim J. 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.

20- Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; 14: 822-827.

21- Guo L, Ma Y, Ward R, Castranova V, Shi X, Qian Y. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clin Cancer Res* 2006; 12: 3344-3354.

22- 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.

23- Zhu ZH, Sun BY, Ma Y, Shao JY, Long H, Zhang X, Fu JH, Zhang LJ, Su XD, Wu QL, Ling P, Chen M, Xie ZM, Hu Y, Rong TH. Three immunomarker support vector machines-based prognostic classifiers for stage IB non-small-cell lung cancer. *J Clin Oncol* 2009; 27: 1091-1099.

24- Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; 356: 11-20.

25- Rami-Porta R, Ball D, Crowley J, Giroux DJ, Jett J, Travis WD, Tsuboi M, Vallières E, Goldstraw P; International Staging Committee; Cancer Research and Biostatistics; Observers to the Committee; Participating Institutions. The IASLC Lung Cancer Staging Project: proposals for the revision of the T descriptors in the forthcoming (seventh) edition of the TNM classification for lung cancer. *J Thorac Oncol* 2007; 2: 593-602.

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

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

28- Martin B, Paesmans M, Mascaux C, Berghmans T, Lothaire P, Meert AP, Lafitte JJ, Sculier JP. Ki-67 expression and patients survival in lung cancer:

systematic review of the literature with meta-analysis. *Br J Cancer* 2004; 91: 2018-2025.

29- Conde E, Suarez-Gauthier A, García-García E, Lopez-Rios F, Lopez-Encuentra A, García-Lujan R, Morente M, Sanchez-Verde L, Sanchez-Cespedes M. Specific pattern of LKB1 and phospho-acetyl-CoA carboxylase protein immunostaining in human normal tissues and lung carcinomas. *Hum Pathol* 2007;38:1351-60.

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

Footnotes - figure legends

Figure 1

Area under the receiver operative characteristic (ROC) curves (AUC) for each Group of Variables taking all groups into account.

Reference **(-------)**. Group A **(--------):** only anatomic extent (TNM) and histological type (AUC: 0.67) variables. Group B **(--------):** clinical variables (AUC: 0.65). Group C **(--------)**: functional and laboratory variables (AUC: 0.57). Group D **(------)**: molecular variables (tissue microarrays) (AUC: 0.65). Integrated Group **(--------)** (Groups A+B+C+D) (AUC: 0.74)

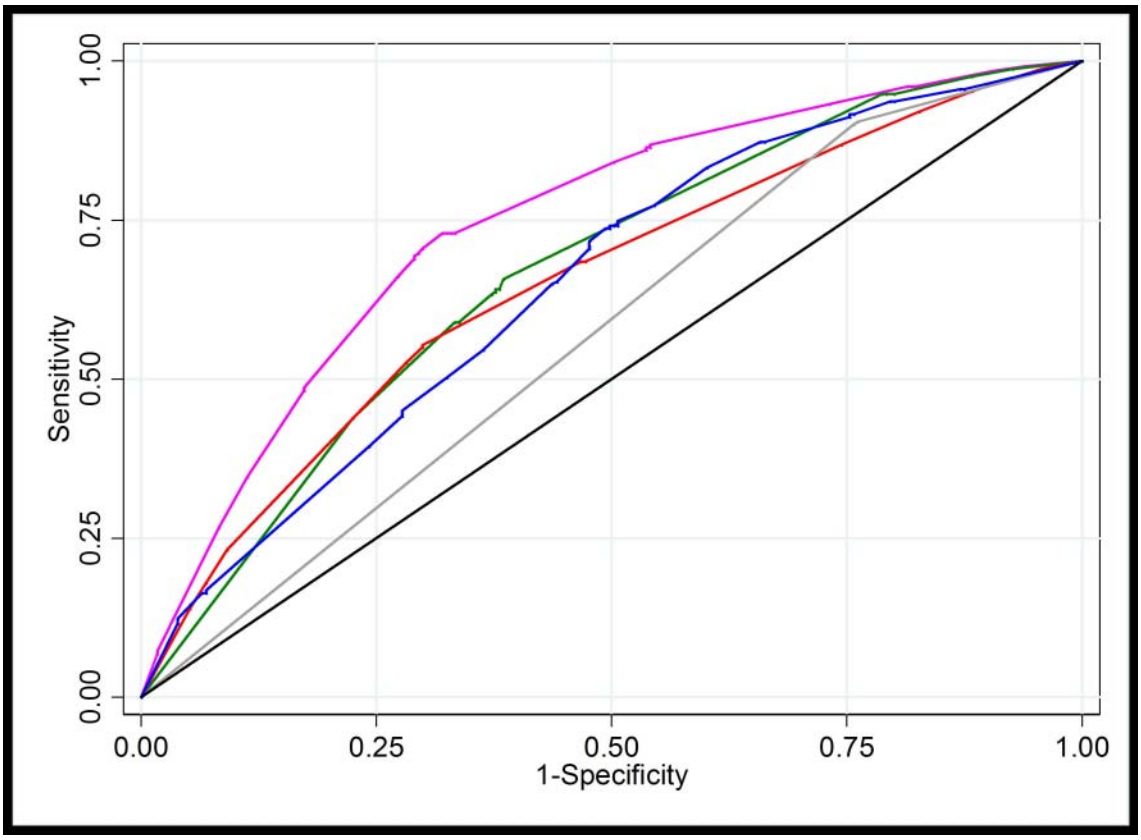


Figure 2

Classification tree. The number of cases over the total number of study cases (n=482) and their probability of overall five year survival (in bold text) is stated for each terminal node. PS: Performance status; T3di: directly invades any of the following: diaphragm, phrenic nerve, mediastinal pleura, pericardium, extrapericardial pulmonary artery or extrapericardial pulmonary vein involvement; Proximal I: Proximal bronchial involvement; Atelectasis: Atelectasis – pneumonitis; ACC: p-Acetyl-CoA-Carboxylase

FIGURE 2. López Encuentra, Angel

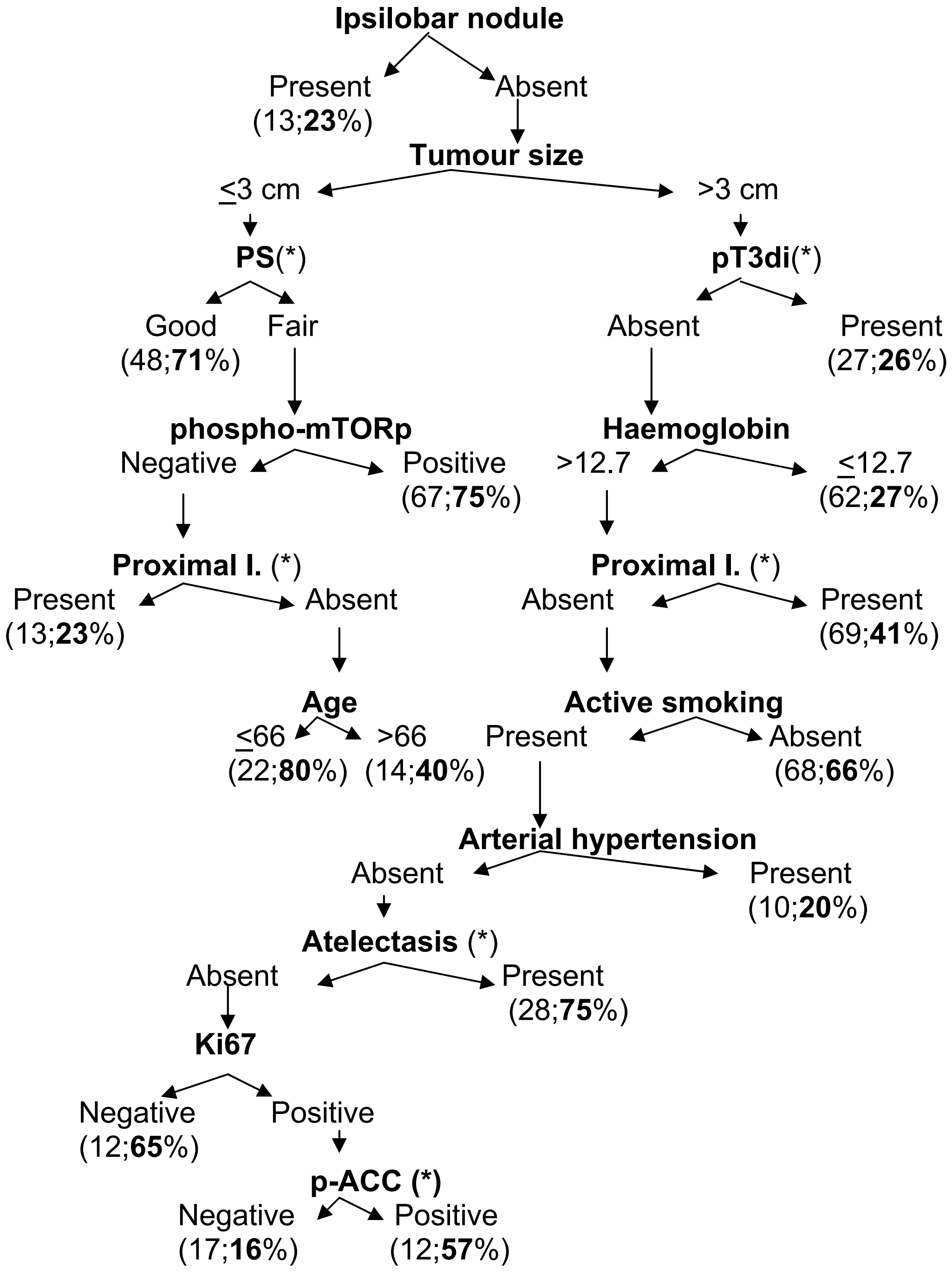
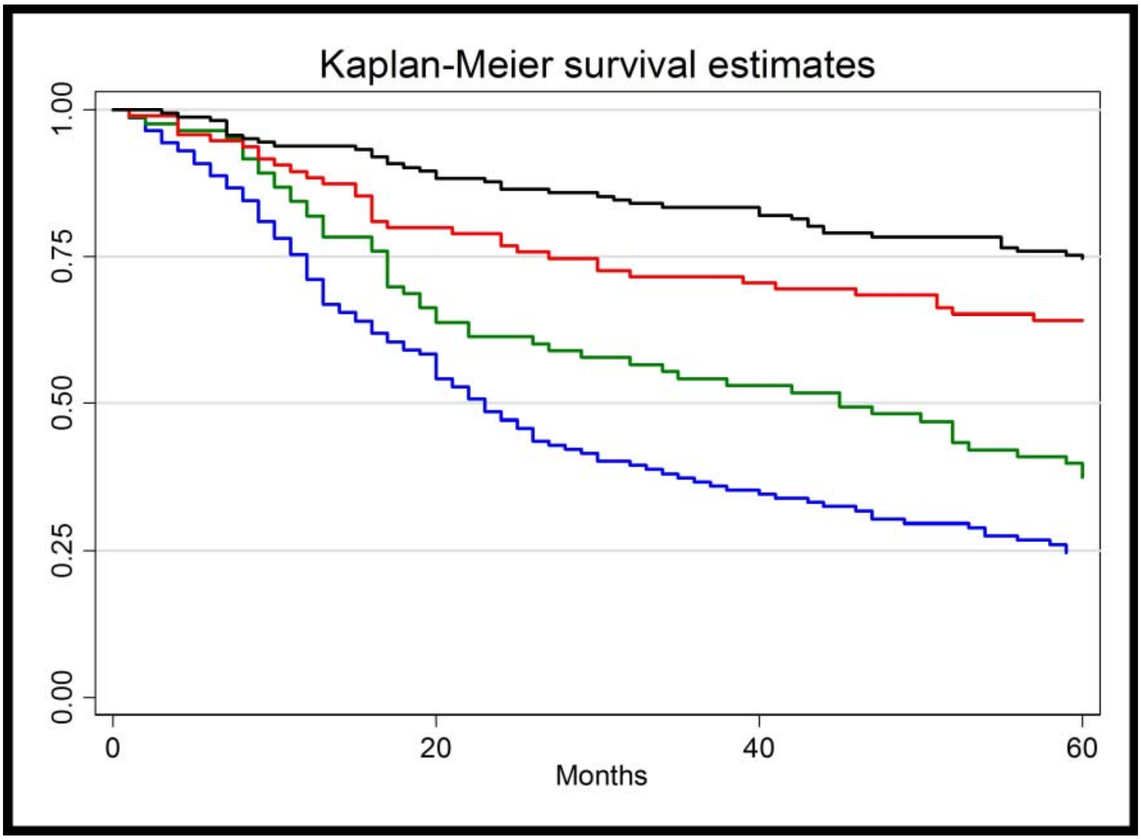


Figure 3

Survival curves of overall survival (5 years) for terminal node groups of similar survival of the classification tree are depicted graphically:

- Group 1: Colour=Black (n=165): terminal node case grouping with overall 5-year survival between 71% and 81% (Figure 2).
- Grupo 2: Colour=red (n=92): terminal node case grouping with overall 5-year survival between 57% and 66% (Figure 2).
- Grupo 3: Colour=green (n=83): terminal node case grouping with overall 5-year survival between 40% and 41% (Figure 2).
- Grupo 4: Colour=blue (n=142): terminal node case grouping with overall survival between 16% and 27% (Figure 2).
-



ACKNOWLEDGMENTS:

Bronchogenic Carcinoma Cooperative Group of the Spanish Society of Pneumology and Thoracic Surgery

General Coordinators:
José Luis Duque-Medina, MD (Hospital Universitario, Valladolid); Angel López-Encuentra, MD (Hospital Universitario 12 de Octubre, Madrid); Ramón Rami-Porta, MD (Hospital Universitario Mutua de Terrassa, Barcelona).

Local representatives:
Julio Astudillo-Pombo, MD (Hospital Germans Trias i Pujol, Barcelona); Josep Maria Jimferrer, MD, Hospital Clinic, Barcelona); Antonio Cantó, MD, (Hospital Clínico, Valencia); Juan Casanova, MD; Manuel Mariñan, MD (Hospital de Cruces, Bilbao); Jorge Cerezal, MD, (Hospital Universitario, Valladolid); Antonio Fernández de Rota, MD; Ricardo Arrabal, MD (Hospital Carlos Haya, Málaga); Federico González Aragoneses, MD; Nicolás Moreno, MD, (Hospital Gregorio Marañón, Madrid); Jorge Freixinet, MD, (Hospital Dr. Negrín , Las Palmas); Nicolás Llobregat, MD, (Hospital Universitario del Aire, Madrid); Nuria Mañes, MD; Helena Hernández, MD, (Fundación Jiménez Díaz, Madrid); Mireia Serra Mitjans, MD, (Hospital Universitario Mutua de Terrassa, Barcelona); José Luis Martín de Nicolás, MD, (Hospital Universitario 12 de Octubre, Madrid); Nuria Novoa (Complejo Hospitalario, Salamanca); Jesús Rodríguez, MD, (Complejo Hospitalario, Oviedo); Antonio José Torres García, MD, (Hospital Universitario San Carlos, Madrid); Mercedes de la Torre (Hospital Juan Canalejo, La Coruña); Abel Sanchez-Palencia, MD, (Hospital Virgen de las Nieves, Granada); Andrés Varela Ugarte, MD; María del Mar Córdoba, MD, (Clínica Puerta de Hierro, Madrid); Yat Wah Pun, MD (Hospital de la Princesa, Madrid).

Data analysis: Agustín Gómez de la Cámara, MD; David Lora Pablos, MSC; Francisco Pozo Rodríguez, MD. Hospital Universitario 12 de Octubre, Madrid. Spain.
Biologic Panel: Montserrat Sánchez-Céspedes, PhD. [*Presently*: Genes and Cancer Group, Programa de Epigenetica y Biologia del Cancer-PEBC, Institut

d’Investigacions Biomediques Bellvitge (IDIBELL), L’Hopitalet de Llobregat, Barcelona, Spain]

Pathologic Panel: **Coordinators:** Esther Conde, MD, Ana Suárez-Gauthier, MD, and Fernando López-Ríos, MD [*Presently:* Laboratorio de Dianas Terapéuticas, Centro Integral Oncológico “Clara Campal”, Hospital Universitario Madrid Sanchinarro, Universidad San Pablo-CEU, Madrid, Spain]

Pathologic Panel: **Participating members**: Guadalupe Renedo, MD, (Fundación Jiménez Díaz, Madrid); Elena García-Lagarto, MD, (Hospital Clínico Universitario, Valladolid); Guadalupe González-Pont, MD, (Hospital Universitario Mutua de Terrasa, Barcelona); Teresa Romero-Saavedra, MD, (Hospital Universitario Dr. Negrín, Las Palmas); José Luis Mate-Sanz, MD, (Hospital Germans Trias i Pujol, Badalona, Barcelona)

We thank Teresa Sotelo, MD, PhD (Hospital Universitario 12 de Octubre, Madrid), for her contributions to this study.