

# SCIENTIFIC REPORTS



OPEN

## Highly Sensitive Marker Panel for Guidance in Lung Cancer Rapid Diagnostic Units

Received: 28 July 2016  
Accepted: 09 December 2016  
Published: 24 January 2017

Sonia Blanco-Prieto<sup>1</sup>, Loretta De Chiara<sup>1</sup>, Mar Rodríguez-Girondo<sup>2,3</sup>, Lorena Vázquez-Iglesias<sup>1</sup>, Francisco Javier Rodríguez-Berrocal<sup>1</sup>, Alberto Fernández-Villar<sup>4</sup>, María Isabel Botana-Rial<sup>4</sup> & María Páez de la Cadena<sup>1</sup>

While evidence for lung cancer screening implementation in Europe is awaited, Rapid Diagnostic Units have been established in many hospitals to accelerate the early diagnosis of lung cancer. We seek to develop an algorithm to detect lung cancer in a symptomatic population attending such unit, based on a sensitive serum marker panel. Serum concentrations of Epidermal Growth Factor, sCD26, Calprotectin, Matrix Metalloproteinases –1, –7, –9, CEA and CYFRA 21.1 were determined in 140 patients with respiratory symptoms (lung cancer and controls with/without benign pathology). Logistic Lasso regression was performed to derive a lung cancer prediction model, and the resulting algorithm was tested in a validation set. A classification rule based on EGF, sCD26, Calprotectin and CEA was established, able to reasonably discriminate lung cancer with 97% sensitivity and 43% specificity in the training set, and 91.7% sensitivity and 45.4% specificity in the validation set. Overall, the panel identified with high sensitivity stage I non-small cell lung cancer (94.7%) and 100% small-cell lung cancers. Our study provides a sensitive 4-marker classification algorithm for lung cancer detection to aid in the management of suspicious lung cancer patients in the context of Rapid Diagnostic Units.

Lung cancer (LC) is the most common cause of cancer-related death worldwide, accounting for 13% of new cancer diagnoses and 19% of total deaths<sup>1</sup>. Despite recent advances in treatment, this neoplasia carries an extremely poor prognosis, with an overall 5-year survival rate of 13%<sup>2</sup>, consequence of the difficulty of detection at an early stage<sup>3</sup>. Therefore, early detection when surgery may be curative is the best way to reduce LC mortality<sup>2,4</sup>.

Opposed to the scenario in U.S., where screening using low-dose computed tomography (CT) among high-risk individuals has been recommended<sup>5</sup>, in Europe there are no LC screening recommendations so far<sup>6</sup>. Among the main reasons are the high rate of false positive results<sup>7–9</sup>, and the awaited upcoming results of the on-going randomized control trials (reviewed in Ruchalski *et al.*<sup>9</sup>).

In Spain, the reduction of the time before diagnosis and staging is a priority<sup>8,10–12</sup>, since both radiological imaging (CT, PET) and invasive procedures for histological confirmation (bronchoscopy, thoracic needle aspiration or thoracentesis) are required in the diagnostic work-up of a LC patient<sup>13,14</sup>. Rapid or Quick Diagnostic Units (RDU/QDU) have been established within the public health system, with the main objective of accelerating the early diagnosis of potentially severe diseases such as cancer, avoiding hospitalisations for purely diagnostic purposes, minimizing hospital-related morbidity, reducing costs, and improving patient satisfaction<sup>15,16</sup>. In Spain approximately 40–50% of the patients attending LC-RDUs display non-cancerous lung pathologies<sup>11,12</sup>. Consequently, clinical decision-making in the setting of LC-RDU would benefit from non-invasive markers that could help predict LC risk in symptomatic individuals, discerning cancerous patients who should be submitted to confirmatory diagnostic procedures from those without cancer in whom a conservative approach could be applied avoiding initially such procedures. Recent reviews have indicated that blood-based markers would be an ideal tool to detect early-stage LC and complement CT imaging<sup>17,18</sup>.

<sup>1</sup>Department of Biochemistry, Genetics and Immunology, Faculty of Biology, Universidad de Vigo. 36310 Vigo, Spain. <sup>2</sup>Department of Medical Statistics and Bioinformatics, Leiden University Medical Center. 2300RC Leiden, The Netherlands. <sup>3</sup>SIDOR Research Group & Centro de Investigaciones Biomédicas (CINBIO), Faculty of Economics and Business Administration, Universidad de Vigo. 36310 Vigo, Spain. <sup>4</sup>Servicio de Neumología Hospital Álvaro Cunqueiro EOVI Vigo, Instituto de Investigación Biomédica de Vigo. 36312 Vigo, Spain. Correspondence and requests for materials should be addressed to M.P.d.I.C. (email: mpaez@uvigo.es)

Marker	Function in Cancer	Usefulness in LC diagnosis
Epidermal Growth Factor (EGF)	Binding of EGF to receptor promotes tumour growth and progression <sup>42</sup>	Suitable discrimination of LC/NSCLC from healthy and benign lung pathologies <sup>19,20</sup>
sCD26	Immune regulation and co-stimulatory activities <sup>43</sup> Tumour suppressor protein in NSCLC <sup>44</sup>	Suitable diagnostic potential for LC vs healthy <sup>21</sup> , healthy/benign <sup>19</sup> or NSCLC-MPE/NSCLC-PMPE <sup>22</sup>
Calprotectin (CAL)	Antimicrobial and pro-inflammatory functions, tumour development <sup>45</sup> Mediate lung metastases <sup>46</sup>	Promising marker in LC derived by pleural effusion <sup>23</sup> and LC <sup>19</sup>
Matrix Metalloproteinases (MMP-1, -7, -9)	Extracellular matrix degradation leading to cancer invasion and metastasis, processing of growth factors, regulation of apoptosis and angiogenesis, tumour-associated inflammation and immune escape <sup>47</sup>	MMP-1: Poor diagnostic capacity for LC vs healthy controls in plasma <sup>48</sup>
		MMP-7: Moderate diagnostic potential for NSCLC vs healthy and benign lung disease <sup>49</sup>
		MMP-9: Good diagnostic potential for LC vs control and benign lung affections <sup>50,51</sup>
Carcinoembryonic Antigen (CEA)	Belongs to immunoglobulin superfamily, acting in cell adhesion and innate immunity <sup>52</sup>	Moderate diagnostic potential for LC vs non-malignant pathologies <sup>4,24,25</sup>
CYFRA 21.1	Fragment of Cytokeratin 19, constituent of cytoskeleton and expressed in epithelial differentiation <sup>53</sup>	Moderate diagnostic potential for LC vs non-malignant pathologies <sup>4,24,25</sup>

**Table 1. Markers Selected for the Development of a Diagnostic Panel for Lung Cancer.** Abbreviations: LC = Lung Cancer, NSCLC = Non Small Cell Lung Cancer, MPE = Malignant Pleural Effusion, PMPE = paramalignant pleural effusion.

We previously described a three-marker panel that included EGF (Epidermal Growth Factor), soluble CD26 (sCD26) and Calprotectin (CAL), showing a considerable discriminatory capacity to detect patients at high risk for LC (83% sensitivity and 87% specificity)<sup>19</sup>. In this new study, together with EGF, sCD26, and CAL, other 5 serum markers were evaluated: MMP-1, -7, -9 (Matrix Metalloprotease -1, -7, -9), CEA (Carcinoembryonic antigen) and CYFRA 21.1, with the aim of improving our previous diagnostic algorithm. These molecules cover a spectrum of biological functions implicated in cancer development and progression, as summarized on Table 1. Since this novel panel is intended to be used in a LC-RDU managed by consultants receiving referrals from primary care doctors, an elevated sensitivity to detect LC among symptomatic patients is imperative.

## Results

**Marker Levels in Lung Cancer and Controls.** Serum levels of the 8 markers analyzed are shown in Table 2, including the median and range for the control group (healthy and benign) and LC. After correction for multiple testing, serum concentrations of EGF, CAL, MMP-1, MMP-7, MMP-9, CEA and CYFRA 21.1 were significantly elevated in LC compared to controls (Mann-Whitney U test,  $P = 0.001$  for EGF, CAL, MMP-9, CEA and CYFRA 21.1;  $P = 0.047$  for MMP-1 and  $P = 0.013$  for MMP-7), while sCD26 levels were notably lower in malignancy relative to controls (Mann-Whitney U test,  $P = 0.001$ ).

All marker levels were found significantly different between healthy controls and cancer subjects (Mann-Whitney U test,  $P = 0.002$  for EGF, sCD26, CAL, MMP-9, CEA and CYFRA 21.1;  $P = 0.018$  for MMP-1 and MMP-7). However, when comparing patients with benign pathologies and cancer, differences in MMP-1 and MMP-7 resulted not significant (Mann-Whitney U test,  $P = 0.448$  for MMP-1 and  $P = 0.090$  for MMP-7). In multivariate linear regression models adjusted for gender, age and smoking status, significant association was again observed for the occurrence of LC and the markers, except for MMP-1 and MMP-7, when considering both the healthy group and all controls. However, only CAL, MMP-9 and CEA maintained significant association with LC regarding benign pathologies.

Furthermore, correlation between the eight markers analysed was also explored using an annotated heatmap (Supplementary Figure S1). Correlations rank from a minimum of 0.038 between EGF and CYFRA 21.1, and a maximum of 0.489 for CAL and MMP-9. Moderate correlations were also observed for several markers, as with EGF with CAL and MMP-9, and the negative correlation of sCD26 with CAL.

The performance of the candidate markers was evaluated by means of ROC curves (Table 2). CAL showed the best potential to discriminate LC from controls (AUC 0.759), followed by CEA (AUC 0.744), CYFRA 21.1 (AUC 0.734) and MMP-9 (AUC 0.729). EGF and sCD26 exhibited AUCs in the range of 0.7, while MMP-1 and MMP-7 demonstrated poor discriminatory capacity (AUC 0.597 and 0.627, respectively).

**Marker Levels by Cancer Histology and Stage.** As displayed in Table 3, LC cases were evaluated based on histology. Statistically significant differences after correction for multiple testing were found between both non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) in relation to controls for sCD26, CEA and CYFRA 21.1 (Mann Whitney U test, NSCLC vs controls:  $P = 0.002$  for sCD26, CEA and CYFRA 21.1; SCLC vs controls:  $P = 0.040$  for sCD26,  $P = 0.002$  for CEA and  $P = 0.012$  for CYFRA 21.1). Levels of EGF, CAL, MMP-7 and MMP-9 resulted different when comparing NSCLC and controls (Mann-Whitney U test,  $P = 0.002$  for EGF,

Marker	Control/Case <sup>a</sup>	Median	Range	P <sup>b</sup>	Adjusted effect (95% CI) <sup>c</sup>	AUC (95% CI)
EGF (pg/mL)	Control	336.86	40.13–1187.06			
	Healthy	247.60	40.13–972.90			
	Benign	419.67	43.25–1187.06			
	LC	571.70	40.75–1716.30	0.001	0.238 (0.140–0.336)*	0.698 (0.615–0.773)
sCD26 (ng/mL)	Control	471.50	122.00–1092.00			
	Healthy	509.00	237.00–886.00			
	Benign	431.50	122.00–1092.00			
	LC	356.00	136.00–1192.00	0.001	−0.097 (−0.148–−0.047)*	0.711 (0.629–0.785)
CAL (ng/mL)	Control	127.87	7.56–421.23			
	Healthy	105.47	7.56–362.29			
	Benign	158.13	33.13–421.23			
	LC	221.21	48.33–438.32	0.001	0.263 (0.190–0.336)*	0.759 (0.679–0.827)
MMP-1 (pg/mL)	Control	5459.61	1186.61–23960.37			
	Healthy	4664.94	1186.61–23635.33			
	Benign	6420.51	1207.70–23960.37			
	LC	7133.87	1450.34–41668.33	0.047	0.089 (−0.008–0.187)	0.597 (0.511–0.679)
MMP-7 (pg/mL)	Control	21761.35	5026.14–79977.27			
	Healthy	21237.20	10383.69–60968.91			
	Benign	22286.74	5026.14–79977.27			
	LC	26710.65	5383.18–79809.13	0.013	0.020 (−0.040–0.081)	0.624 (0.539–0.705)
MMP-9 (ng/mL)	Control	177.60	52.79–3611.59			
	Healthy	177.19	52.79–743.25			
	Benign	181.74	57.41–3611.59			
	LC	311.86	21.06–1914.00	0.001	0.270 (0.167–0.373)*	0.729 (0.648–0.801)
CEA (pg/mL)	Control	837.26	170.84–4070.71			
	Healthy	929.50	171.03–4070.71			
	Benign	559.51	170.84–2828.15			
	LC	2051.64	141.16–136039.19	0.001	0.494 (0.333–0.656)*	0.744 (0.663–0.814)
CYFRA 21.1 (pg/mL)	Control	227.05	0.00–19314.33			
	Healthy	0.00	0.00–16592.63			
	Benign	1052.47	0.00–19314.33			
	LC	3007.09	0.00–173410.17	0.001	1.080 (0.587–1.573)*	0.734 (0.653–0.805)

**Table 2. Serum Markers in Lung Cancer and Controls in the Training Set.** Abbreviations: LC = Lung Cancer. <sup>a</sup>Sample size in training set: Control n = 72 (Healthy n = 36, Benign n = 36), LC n = 68. <sup>b</sup>Mann-Whitney U test for comparison between the cancer and control group corrected by Benjamini-Hochberg method to control familywise error under multiple comparisons. <sup>c</sup>Adjusted effects and 95% confidence intervals of the case/control status on each of the log-transformed markers considered as outcome in lineal regression model adjusted for gender, age and smoking. \*P-value < 0.001.

CAL and MMP-9,  $P = 0.048$  for MMP-7), but not for SCLC (Mann-Whitney U test,  $P = 0.798$  for EGF and MMP-9,  $P = 0.112$  for CAL and  $P = 0.056$  for MMP-7).

The potential of the markers to detect early stage LC was analysed according to tumour stage (Table 3). EGF, CAL and MMP-9 were the only molecules significantly altered after multiple testing correction in NSCLC stage I + II (Mann-Whitney U test,  $P = 0.002$  for EGF and MMP-9,  $P = 0.017$  for CAL), suggesting their usefulness for diagnosis at earliest stages. For late stage NSCLC, all markers except for MMP-1 displayed significant differences (Mann-Whitney U test,  $P = 0.002$  for EGF, sCD26, CAL, MMP-9, CEA and CYFRA 21.1,  $P = 0.023$  for MMP-7). Regarding SCLC a dramatic reduction of EGF, CAL and MMP-9 levels in limited SCLC prevents the distinction from non-cancer patients. After adjusting for the common risk factors gender, age and smoking, the same associations were maintained, with the exception of SCLC association with sCD26, and the lack of significance with either LC histology or NSCLC stages and MMP-7.

**Association between Clinical Parameters and Marker Levels.** The association of marker concentrations with clinical variables gender, age and smoking status is presented in Supplementary Table S1. sCD26 levels were significantly higher in women relative to men (Mann-Whitney U test,  $P = 0.018$ ), with no other marker influenced by gender. Older age was associated significantly with higher levels of MMP-7 and CYFRA 21.1 (Mann-Whitney U test,  $P < 0.001$  and  $P = 0.004$ , respectively), whereas sCD26 levels diminished with age. Patients with smoking habits had significantly increased serum concentrations of EGF and MMP-7 in relation to never smokers (Mann-Whitney U test,  $P = 0.028$  for EGF and  $P = 0.026$  for MMP-7).

**Multimarker Panel and Classification Algorithm for Lung Cancer.** Lasso regression was employed to simultaneously derive a multivariate panel of markers and an optimal cut-off for LC, with the criterion of maximizing specificity for a predefined sensitivity of 95%. The resulting classification rule as well as the optimal Lasso penalization parameter are available in Supplementary Material S1 and Supplementary Figure S2, respectively. Additionally, Supplementary Table S2 includes several diagnostic measurements for the proposed model.

Variability in the proportion of males and smoking, and differences in age between LC and controls, as well as influence of these variables on marker levels motivated their inclusion in the model. Application of Lasso procedure on the training set led to the establishment of a 4-marker panel composed of EGF, sCD26, CAL and CEA. A clinical model composed of gender, age and smoking was also established by logistic regression for comparison.

Performance and ROC curves of this marker panel and single markers for LC diagnosis, besides the clinical model, are presented on Table 4 and Fig. 1, respectively. The 4-marker panel demonstrated a good discriminatory capacity to differentiate LC patients from controls with an AUC of 0.873, showing 97% sensitivity and 43% specificity for LC detection corresponding to a 0.266 cut-off. This combination of markers outperforms the individual markers in terms of specificity. In relation to the clinical model, a lower discriminatory ability was displayed as compared to our proposed multivariate panel (AUC = 0.717 (0.637–0.799), DeLong test  $P$ -value < 0.0001). At the desired sensitivity of 95%, the decision rule based on such clinical model renders a poor specificity of 26%. Based on our model and assuming a prevalence of LC of 44.4%, corresponding to the RDU of the Pneumology Service of EOXI Vigo, an optimal Negative Predictive Value (NPV) of 94.7% was reached, and a moderate Positive Predictive Value (PPV) of 57.6%.

To further verify the performance of the 4-marker panel for prediction of LC, the resulting classification algorithm developed in the training set was tested in an independent validation set. Descriptive statistics for each marker of the panel are given in Supplementary Table S3 according to histology and stage. In the validation set the marker panel showed an AUC of 0.837, with a sensitivity of 91.7% and a moderately higher specificity of 45.4%, based on the 0.266 cut-off established (Table 4). Regarding the clinical model, the inferior discriminatory capacity was again evidenced by the AUC of 0.659 (0.488–0.816) (DeLong test  $P$ -value = 0.0003).

**Classification Accuracy of the 4-Marker Panel for Lung Cancer and Control Subgroups.** To deeply assess the performance of our classification rule we examined its ability to correctly classify specific subgroups of LC patients and controls (Table 5). Training and validation populations were combined, and sensitivity for the histological subgroups and stage was calculated at fixed 43% specificity (0.266 cut-off). The classification rule identified with high sensitivity stage I NSCLC (94.7%) and stage II (100%), similarly to advanced stages III and IV (95.2 and 94.6%, respectively). The most prevalent NSCLC type, adenocarcinoma (ADC), also demonstrated a high sensitivity (93.2%), as in Squamous Cell Carcinoma (SqCC) and Large Cell Carcinoma (LCC) (100% both). All patients with SCLC were likewise detected with 100% sensitivity.

Among non-cancerous patients the panel correctly classified 41 out of 94 controls (43.6%), yielding a specificity of 40.9% for healthy and 46% for benign conditions of the lung.

Table 5 also includes the classification accuracy based on the results of the CT scan, specifically when no mass was detected. When additionally no nodules were found, the panel correctly classified all LC cases (7/7; 100% sensitivity). On the contrary, in the presence of nodules, our panel was able to classify 1 out of 3 controls (33.3% specificity).

## Discussion

Classification algorithms capable of guiding clinical decision-making constitute a valuable tool that can help predict LC, besides complement CT imaging<sup>17,18</sup>. In a previous work we described a three-marker panel for high-risk patients including the molecules EGF, sCD26 and CAL, and gender and age as confounders, and their implication in lung carcinogenesis was enclosed<sup>19–23</sup>. Here we provide an improved classification algorithm achieving a superior sensitivity for LC in the context of RDU. In this refined algorithm, besides the smoking status, the routinely used CEA was incorporated, corroborating its diagnostic capacity especially for late-stage tumours<sup>4,24–26</sup>. Briefly, our approach involves the measurement of EGF, sCD26, CAL and CEA to generate a classification score for each individual to predict LC.

As for colorectal and breast cancer, LC could also benefit from screening programs. However, at this time in Europe there are no LC screening recommendations though The European Society of Radiology and the European Respiratory Society recommend screening within a clinical trial or in routine clinical practice at certified medical centres<sup>6</sup>. Instead, the strategy implemented in many European hospitals to achieve an early detection is the acceleration in the time to diagnosis in the so called Rapid Diagnostic Units for LC<sup>10–12,15,16</sup>. Consequently, we intended to design a marker-based classification algorithm to be used in these Units, where the priority is to detect all LC cases (high sensitivity), in order to select those patients that should be immediately submitted to more invasive tests.

Individual analyses evidenced the usefulness of EGF, sCD26, CAL and CEA among the 8 molecules assayed, with AUCs between 0.698–0.759 for the training set and 0.716–0.871 for the validation set, headed in both cohorts by CAL. Among the four markers, differences were more frequent comparing NSCLC and controls, even at early stages as in the case of EGF and CAL. In relation to SCLC, sCD26 and CEA were the markers that better differentiated this histological group. The individual diagnostic potential of the four markers resulted in a modestly specific signature for the detection of LC when combined through a multivariate logistic Lasso regression approach that provided, by design, desirable sensitivity. This strategy demonstrated 97% sensitivity in the training set and for a >0.266 cut-off the classification algorithm showed a specificity of 43%. In the validation set sensitivity resulted in a fine 91.7% and 45.4% specificity. This modest specificity is of value in the clinical context of RDU with patients with respiratory symptoms and/or LC suspicion. Performance of our marker panel also outperforms that of a clinical model constituted by gender, age and smoking.

Marker	Control/Case <sup>a</sup>	Median	Range	P <sup>b</sup>	Adjusted effect (95% CI) <sup>c</sup>
<b>EGF</b> (pg/mL)	<b>Control</b>	<b>336.86</b>	<b>40.13–1187.06</b>		
	<b>NSCLC</b>	<b>601.97</b>	<b>144.23–1176.15</b>	<b>0.002</b>	<b>0.277 (0.168–0.386)*</b>
	Early (I + II)	785.28	388.23–1159.73	0.002	0.409 (0.236–0.581)*
	Late (III + IV)	528.51	144.23–1176.15	0.002	0.231 (0.109–0.354)*
	<b>SCLC</b>	<b>295.48</b>	<b>40.75–1716.30</b>	<b>0.798</b>	<b>−0.039 (−0.293–0.216)</b>
	Limited	55.97	40.75–201.63	—	
	Extended	440.64	264.30–1716.30	—	
<b>sCD26</b> (ng/mL)	<b>Control</b>	<b>471.50</b>	<b>122.00–1092.00</b>		
	<b>NSCLC</b>	<b>356.00</b>	<b>136.00–945.00</b>	<b>0.002</b>	<b>−0.095 (−0.153– −0.038)*</b>
	Early (I + II)	432.00	206.00–640.00	0.235	−0.032 (−0.114–0.049)
	Late (III + IV)	341.00	136.00–945.00	0.002	−0.116 (−0.179– −0.054)*
	<b>SCLC</b>	<b>294.00</b>	<b>208.00–1092.00</b>	<b>0.040</b>	<b>−0.073 (−0.184–0.037)</b>
	Limited	370.00	339.00–1192.00	—	
	Extended	288.00	208.00–541.00	—	
<b>CAL</b> (ng/mL)	<b>Control</b>	<b>127.87</b>	<b>7.56–421.23</b>		
	<b>NSCLC</b>	<b>221.06</b>	<b>87.19–438.32</b>	<b>0.002</b>	<b>0.250 (0.160–0.341)*</b>
	Early (I + II)	193.72	120.73–340.67	0.017	0.188 (0.033–0.343)*
	Late (III + IV)	238.59	87.19–438.32	0.002	0.275 (0.172–0.378)*
	<b>SCLC</b>	<b>245.69</b>	<b>48.33–422.20</b>	<b>0.112</b>	<b>0.203 (−0.014–0.419)</b>
	Limited	97.32	91.06–422.20	—	
	Extended	279.34	48.33–355.19	—	
<b>MMP-1</b> (pg/mL)	<b>Control</b>	<b>5459.61</b>	<b>1186.61–23960.37</b>		
	<b>NSCLC</b>	<b>7132.52</b>	<b>1450.34–41668.33</b>	<b>0.093</b>	<b>0.071 (−0.049–0.192)</b>
	Early (I + II)	7800.42	1784.41–30180.33	0.235	0.087 (−0.106–0.279)
	Late (III + IV)	7132.52	1450.34–41668.33	0.131	0.066 (−0.065–0.198)
	<b>SCLC</b>	<b>7135.23</b>	<b>2220.44–18900.42</b>	<b>0.279</b>	<b>0.083 (−0.158–0.323)</b>
	Limited	7135.23	4890.35–18900.42	—	
	Extended	8114.52	2220.44–14246.29	—	
<b>MMP-7</b> (pg/mL)	<b>Control</b>	<b>21761.35</b>	<b>5026.14–79977.27</b>		
	<b>NSCLC</b>	<b>26485.45</b>	<b>5383.18–79809.13</b>	<b>0.048</b>	<b>0.021 (−0.052–0.095)</b>
	Early (I + II)	24615.56	8551.00–51000.43	0.726	−0.031 (−0.136–0.074)
	Late (III + IV)	26680.26	5383.18–79809.13	0.023	0.039 (−0.041–0.118)
	<b>SCLC</b>	<b>28177.70</b>	<b>8231.19–43429.13</b>	<b>0.056</b>	<b>0.052 (−0.082–0.186)</b>
	Limited	26874.44	26710.65–40252.39	—	
	Extended	30110.14	8231.19–43429.13	—	
<b>MMP-9</b> (ng/mL)	<b>Control</b>	<b>177.60</b>	<b>52.79–3611.59</b>		
	<b>NSCLC</b>	<b>340.19</b>	<b>21.06–1914.00</b>	<b>0.002</b>	<b>0.257 (0.135–0.379)*</b>
	Early (I + II)	379.94	174.70–1688.00	0.002	0.285 (0.101–0.470)*
	Late (III + IV)	299.65	21.06–1914.00	0.002	0.244 (0.106–0.382)*
	<b>SCLC</b>	<b>225.61</b>	<b>65.58–786.34</b>	<b>0.798</b>	<b>0.013 (−0.235–0.262)</b>
	Limited	115.93	65.58–253.19	—	
	Extended	298.21	76.78–786.34	—	
<b>CEA</b> (pg/mL)	<b>Control</b>	<b>837.26</b>	<b>170.84–4070.71</b>		
	<b>NSCLC</b>	<b>1783.93</b>	<b>141.16–136039.19</b>	<b>0.002</b>	<b>0.467 (0.266–0.668)*</b>
	Early (I + II)	1093.08	353.26–21684.85	0.140	0.154 (−0.050–0.358)
	Late (III + IV)	2750.49	141.16–136039.19	0.002	0.567 (0.348–0.787)*
	<b>SCLC</b>	<b>3704.96</b>	<b>1147.58–82300.26</b>	<b>0.002</b>	<b>0.818 (0.541–1.094)*</b>
	Limited	2092.47	1147.58–82300.26	—	
	Extended	3884.15	1667.40–29844.72	—	
<b>CYFRA 21.1</b> (pg/mL)	<b>Control</b>	<b>227.05</b>	<b>0.00–19314.33</b>		
	<b>NSCLC</b>	<b>2910.66</b>	<b>0.00–173410.17</b>	<b>0.002</b>	<b>0.939 (0.348–1.531)*</b>
	Early (I + II)	1181.22	0.00–7309.16	0.104	0.503 (−0.382–1.387)
	Late (III + IV)	4791.34	0.00–173410.17	0.002	1.105 (0.459–1.752)*
	<b>SCLC</b>	<b>5886.75</b>	<b>0.00–12228.10</b>	<b>0.012</b>	<b>1.285 (0.154–2.415)*</b>
	Limited	1754.30	219.64–6610.22	—	
	Extended	5965.40	0.00–12228.10	—	

**Table 3. Distribution of Serum Markers in Lung Cancer by Histology and Stage and Comparison with Controls in the Training Set.** Abbreviations: NSCLC = Non-Small Cell Lung Cancer, SCLC = Small Cell Lung Cancer. <sup>a</sup>Sample size in training set: Control n = 72, NSCLC n = 59 (Early stage n = 16, Late stage n = 43), SCLC n = 9 (Limited stage n = 3, Extended stage n = 6). <sup>b</sup>Mann-Whitney U test for comparison between the control groups and lung cancer stratified by histology and stage corrected by Benjamini-Hochberg method to control

familywise error under multiple comparisons. <sup>c</sup>Adjusted effects and 95% confidence intervals of histology and stage on each of the log-transformed markers considered as outcome in lineal regression model adjusted for gender, age and smoking. \**P*-value statistically significant.

Training Set	Cut-off	Sn (%)	Sp (%)	PPV <sup>a</sup> (%)	NPV <sup>a</sup> (%)	AUC (95% CI) <sup>b</sup>
Multivariate Algorithm: EGF, sCD26, CAL, CEA	>0.266	97	43	57.6	94.7	0.873 (0.811–0.925)
EGF	>178.48 pg/mL	95	22.2	49.4	84.8	0.698 (0.615–0.773)
sCD26	≤637.2 ng/mL	95	13.9	46.8	77.7	0.711 (0.629–0.785)
CAL	>96.37 ng/mL	95	30.6	52.2	88.4	0.759 (0.679–0.827)
CEA	>258.2 pg/mL	95	11.1	46	73.6	0.744 (0.663–0.814)
Clinical Model <sup>c</sup>	>0.237	95	26.4	50.8	86.9	0.717 (0.637–0.799)
<b>Validation Set</b>						
Multivariate Algorithm: EGF, sCD26, CAL, CEA	>0.266	91.7	45.4	57.3	87.3	0.837 (0.718–0.936)
Clinical Model <sup>c</sup>	>0.237	91.7	27.3	50.2	80.5	0.659 (0.488–0.816)

**Table 4. Performance of the Four-Marker Panel and EGF, sCD26, CAL and CEA in the Diagnosis of Lung Cancer.** Abbreviations: Sn = Sensitivity, Sp = Specificity, PPV = Predictive Positive Value, NPV = Negative Predictive Value. <sup>a</sup>Positive and negative predictive values were estimated assuming a prevalence of lung cancer of 44.4% (QDU of the Pneumology Service of Hospital Álvaro Cunqueiro EOXI Vigo). <sup>b</sup>AUC and 95% CI evaluated in the training test is not protected against overfitting. <sup>c</sup>Clinical model includes gender, age and smoking.

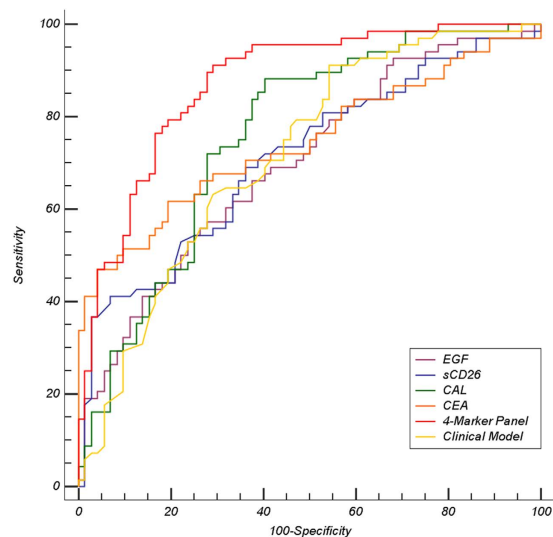
The classification accuracy including training and validation cohorts showed an overall sensitivity of 95.6% for LC. Among the 95% of NSCLC patients correctly classified, 94.7% of stage I tumours were detected. Regarding SCLC, the classification algorithm was effective for all the cases. Among controls, overall specificity resulted 43.6% and was not greatly affected by the nature of the controls themselves. Given the clinical dilemma of indeterminate nodules detected on CT-based screening due to elevated false positive rates<sup>8</sup>, we also evaluated our algorithm according to the absence/presence of nodules. All LC cases (100%) that had a negative CT-scan were correctly classified (6 out of 6 NSCLC and the SCLC case). On the contrary, among controls bearing nodules, our panel classified correctly 1 out of 3 patients. It should be noted that CT-scan data was available for all LC cases but only for 13.8% of controls (3 healthy and 10 benign cases), limiting the analysis.

In the last years several diagnostic multianalyte panels have been proposed for LC, with variable criteria for patient selection such as inclusion limited to NSCLC or absence of controls bearing benign pulmonary pathologies. Studies comparable to ours, at least with similar study population, are scarce. Molina *et al.*<sup>25</sup> proposed a six marker panel (CEA, CA 15.3, Squamous Cell Carcinoma Antigen –SCC–, CYFRA 21.1, Neuron Specific Enolase –NSE– and Progastrin-releasing Peptide –ProGrp–) for patients with suspected LC based on the criterion of any of the markers elevated, proving a sensitivity of 88.5% and specificity of 82%, not validated.

Other studies document protein models combined with CT imaging techniques. Yang *et al.*<sup>24</sup> reported for high-risk patients with no lesions on CT scan a panel, which resulted positive when at least one of the markers CEA, SCC, CYFRA 21.1 and Progastrin-releasing Peptide was altered, yielding a sensitivity of 76.6% and specificity of 94.4%, though they do not report data on another independent sample set. The algorithm established by Patz *et al.*<sup>27</sup> based on the combination of nodule size and CEA, alpha-antitrypsin and SCC, rendered acceptable performance for classifying patients with indeterminate nodules (92% sensitivity and 74% specificity).

To date, only two blood tests based on marker panels have been translated into clinical or commercial setting. The EarlyCTD-Lung, which measures autoantibodies, was developed for the early detection of LC in high-risk population or as adjunct to CT<sup>28</sup>. Its performance was demonstrated in clinical practice, yielding 41% sensitivity and 87% specificity. The PAULA's test (Protein Assays Using Lung cancer Analytes) is a 4-marker panel comprising three tumour antigens (CEA, CA125 and CYFRA 21.1) and one autoantibody (NY-ESO1), intended for early NSCLC tumours in high-risk patients. In a validation set the panel discriminated NSCLC (with 67% early-stage) from healthy controls with a sensitivity of 77% and specificity of 80%. However its clinical applicability is limited since benign conditions were not included<sup>4</sup>. None of the cited studies pursued such a high sensitivity as we do, which would probably derive in a diminished specificity. In these circumstances, we would affirm the promising value of our 4-marker panel.

Our model building procedure is based on regularized regression models which are intended to be more flexible and resistant to overfitting compared to stepwise approaches<sup>29–31</sup>. Furthermore, by design, our method identifies models which guarantee the optimization of the derived classification rule by choosing the penalty parameter and cut-off which maximizes specificity, assuring a predefined sensitivity. The adaptive nature of our method constitutes one of the strengths of our study. Alternative model building techniques established by first choosing a logistic model based on a reduced set of variables by minimizing the AIC or BIC, and then determining a cut-off depending on the given classification setting, are less flexible, and in general our method outperforms these approaches since it is specifically designed for optimizing classification performance. Moreover, approaches based on exhaustive evaluation of all possible sub-models become rapidly unfeasible when increasing



**Figure 1. ROC Curve Analysis for Lung Cancer Prediction in the Training Set.** ROC curves are shown for each individual marker included in the classification algorithm, together with the clinical model and the 4-marker panel derived from logistic Lasso regression. Training set included 68 lung cancer cases and 72 controls (36 healthy and 36 benign respiratory pathologies).

the number of candidate markers, whereas our approach, since it relies on shrinkage, is expected to perform well in such situations.

In Supplementary Table S4 we have included two logistic regression models (two-stage) and derived classification rules for selected 90% and 95% sensitivities. As observed, the obtained models and classification rules are not uniformly optimal and their performance varies according to the classification situation. For example, the classification rule based on BIC performs well for the cut-off that provides 95% sensitivity, while its performance is considerably worse for the cut-off corresponding to 90% sensitivity. Alternatively, if we focus on the AIC as optimal criterion, this method outperforms alternative Lasso-based and BIC for 90% sensitivity, but it presents an inferior performance when we focus on higher sensitivities.

Given the complex challenge of developing an optimal diagnostic panel for LC, a proper study design is also of crucial importance. Besides the consciousness in the statistical approach aforementioned, the inclusion of both benign and healthy individuals in the control group, as well as the two main histological tumours (NSCLC and SCLC) is a strong point of our study. Another important feature is that samples from all the individuals were prospectively collected at their first visit to the Pneumology Service in the presence of respiratory symptoms, reflecting the clinical setting of a RDU. For the refinement of the diagnostic algorithm we have also included information related to tobacco, which constitutes a well-established risk factor<sup>32</sup> and is usually not contemplated in studies developing diagnostic panels.

One of the advantages of our classification algorithm is that only 4 molecules comprise the panel, and 2 of them are already established in hospitals: CEA is routinely measured for various types of tumours, while CAL is also quantified for its utility in inflammatory colon processes<sup>33</sup>. This makes our 4-marker panel simple and affordable to guide clinical decision-making and complement CT scan. Additionally, we are currently working on an interactive web application to facilitate the implementation of the classification algorithm in the biomedical community, based on the Shiny web application for R<sup>34</sup>.

Regarding the limitations of the study, the number of patients was modest for both training and validation sets, particularly for SCLC cases. A possible shortcoming could be the lack of information related to tobacco consumption, which perhaps could have contributed to the improvement of the diagnostic algorithm.

In summary, we defined a modestly specific 4-marker classification algorithm that provided, by design, desirable sensitivity for the detection of LC, conceived to be useful among symptomatic high-risk individuals derived to LC-RDU. The next step along the complicated road to reach the clinical implementation is the validation of our panel in a large, multi-centric cohort.

## Methods

**Study Population.** Between May 2007 and January 2011, 186 patients with respiratory symptoms were prospectively recruited at the Pneumology Service of Hospital Álvaro Cunqueiro EOXI Vigo (Spain). The study population included patients finally diagnosed of LC, and a control cohort with subjects diagnosed of benign lung disease and healthy subjects with no respiratory pathology. Exclusion criteria included relapse or progression of a cancer previously diagnosed, and chemo- or radiotherapy treatment.

Clinical guidelines from the American College of Chest Physicians were followed for LC diagnosis<sup>13,14</sup>. Histological assessment of tumours followed the WHO criteria<sup>35</sup> and staging was performed according to the 7<sup>th</sup> edition of TNM<sup>36</sup>.

		Cases correctly classified <sup>a</sup> /Total cases	% Sn at 43% Sp
<b>Lung Cancer</b>		88/92	95.6
NSCLC		76/80	95
	I	18/19	94.7
	II	3/3	100
	III	20/21	95.2
	IV	35/37	94.6
	ADC	41/44	93.2
	SqCC	20/20	100
	LCC	13/13	100
	BAC	1/2	50
	ND	1/1	100
SCLC		12/12	100
		Cases correctly classified <sup>a</sup> /Total cases	% Sp at 95% Sn
<b>Control</b>		41/94	43.6
Healthy		18/44	40.9
Benign		23/50	46
	RI	19/41	46.3
	ILD	4/9	44.4
<b>Accuracy in the Classification of Cancer and Controls regarding CT imaging</b>			
Absence of Nodules in CT scan		Cases correctly classified <sup>a</sup> /Total cases	% Sn at 43% Sp
<b>Lung Cancer</b>		7/7	100
NSCLC		6/6	100
	I	1/1	100
	II	2/2	100
	III	1/1	100
	IV	2/2	100
SCLC	Extended	1/1	100
Presence of Nodules in CT scan		Cases correctly classified <sup>a</sup> /Total cases	% Sp at 95% Sn
<b>Control</b>		1/3	33.3
Healthy		0/1	0
Benign		1/2	50

**Table 5. Classification Accuracy of the Multivariate Algorithm for Subgroups of Patients in the Combined Set (Training and Validation Set).** Abbreviations: Sn = Sensitivity, Sp = Specificity, NSCLC = Non Small Cell Lung Cancer, ADC = Adenocarcinoma, SqCC = Squamous Cell Carcinoma, LCC = Large Cell Carcinoma, BAC = Bronchioloalveolar Carcinoma, ND = Not Differentiated Carcinoma, SCLC = Small Cell Lung Cancer, RI = Respiratory Infection, ILD = Interstitial Lung Disease. <sup>a</sup>Cut-off  $p = 0.266$  for a sensitivity of 97% and specificity of 43% in the training set.

Recruited individuals were divided into a training set for panel development, and in another set for validation of the algorithm. The training set consisted of 140 individuals and included 68 LC cases (80.9% men, median age 69.5 years). The control cohort included 72 subjects with a median age of 61 years and 63.6% males. The validation set consisted of 46 individuals (24 LC and 22 controls). Patient demographics are outlined in Supplementary Table S5.

The study followed the clinical-ethical practices of the Spanish Government and the Helsinki Declaration, and was approved by the Galician Ethical Committee for Clinical Research. All patients provided written informed consent.

**Determination of Markers Concentration.** Blood samples were collected from all patients at their first visit to the Service, when bronchoscopy was performed. Serum was obtained and stored at  $-20^{\circ}\text{C}$  until analysis.

Measurement of EGF (R&D Systems, Minneapolis, USA), sCD26 (eBioscience, Wien, Austria) and Calprotectin (Hycult Biotechnology, Uden, the Netherlands) concentrations were conducted using enzyme-linked immunosorbent assays (ELISA). Absorbance readings were collected on an EnVision Multilabel Plate Reader (Perkin Elmer).

To measure the amount of serum MMPs, CEA and CYFRA 21.1 multiplexed bead-based immunoassays were used. Levels of MMP-1, MMP-7 and MMP-9 were determined with the Human MMP Panel 2 Magnetic Bead Kit, while CEA and CYFRA 21.1 were part of the Circulating Cancer Biomarker Magnetic Bead Panel 1 (EMD Millipore, Missouri, USA). Fluorescence was read on a Luminex 200™ with BioPlex Manager™ software (Bio-Rad, Hercules, CA), using a 5-parameter logistic fitting for deriving protein concentration in samples.



**Statistical Methods.** *Individual Marker Analysis.* Non-parametric Mann-Whitney U test was used for two-sample group comparisons of continuous variables, while Fisher test was applied for comparison of qualitative variables. Benjamini-Hochberg method for controlling the false discovery rate<sup>37</sup> was used to correct *P*-values for multiple group comparisons. Linear regression models were used to study markers' association with LC presence, histology and stage adjusted for the risk factors gender, age and smoking. The discriminatory ability of markers for LC was evaluated by Receiver Operating Characteristics Curve (ROC) providing the Area Under the Curve (AUC). All tests were two-sided and *P*-values  $\leq 0.05$  were considered statistically significant. Statistical software SPSS 22.0 (SPSS Inc., Chicago, IL) and R program package (Wirtschafts Universität, Wien, Austria) were used to perform these analyses.

**Marker Panel Selection and Classification Algorithm Development.** Marker concentrations were  $\log_{10}$ -transformed before multivariate analysis to reduce the skewness. We derived a classification rule based on a multivariate combination of the studied markers based on logistic Lasso regression<sup>38</sup> fitted in the training set and including age, gender and smoking as fixed effects. This procedure was also used to obtain a clinical model in which only the variables age, gender and smoking were included. Lasso regularization imposes a penalization over the maximum likelihood estimates of the usual regression coefficients so that they are shrunk towards zero. Actually, some of the resulting coefficients can be exactly zero, and hence Lasso shrinkage performs automatic variable selection. The optimal amount of shrinkage is controlled by the selection of the penalization parameter which maximizes the out-of-sample performance (in terms of some pre-defined criterion) of the model. In our algorithm, we simultaneously chose the penalty parameter and cut-off point which provides the classification rule with maximum specificity, given a fixed value of sensitivity (95%). Namely, we use 10-fold cross validation in the training set and for each possible value of the penalty parameter we apply the resulting estimated coefficients to the out-of-sample data, obtaining case probability scores for each observation of the training set. Each of these scores was subsequently dichotomized to guarantee the desired level of sensitivity. Finally, we choose the penalty parameter which maximized the specificity. Further details concerning the Lasso procedure and the proposed algorithm are displayed as Supplementary Material S1.

For prediction of a new individual's diagnosis, the selected classification rule was applied. Based on the coefficients of the regression model, the classification algorithm calculates for a new patient a single score based on the estimated predicted probabilities (*p*) of presenting lung cancer as a function of markers concentrations and demographic variables. A new individual will be classified as cancer if *p* is higher than the cut-off estimated in the training set, while classified as non-cancer when the resulting score is below the cut-off.

Applying the Lasso regression model to the train and test samples, their probability scores were obtained and the diagnostic performance of the classification rule was evaluated by providing sensitivity, specificity and predictive values. ROC curves were elaborated for both the Lasso-based marker model and clinical model, providing the AUC. DeLong test was applied for comparison of AUC values of these models<sup>39</sup>.

For sake of comparison, we evaluated the performance of two alternative two-stage methods based on first selecting an optimal logistic model, based, respectively, on exhaustive sub-model evaluation and selection based on minimization of Akaike information criterion (AIC)<sup>40</sup> and Bayesian information criterion (BIC)<sup>41</sup> and secondly, determining the optimal cut-off for guarantying the desired level of sensitivity.

All multivariate calculations were performed using the R program package (Wirtschafts Universität, Wien, Austria).

## References

- Torre, L. A. *et al.* Global cancer statistics, 2012. *C.A. Cancer J. Clin.* **65**, 87–108 (2015).
- Francisci, S. *et al.* Survival patterns in lung and pleural cancer in Europe 1999–2007: Results from the EUROCARE-5 study. *Eur. J. Cancer.* **51**, 2242–2253 (2015).
- Walters, S. *et al.* Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. *Thorax.* **68**, 551–564 (2013).
- Doseeva, V., Colpitts, T., Gao, G., Woodcock, J. & Knezevic, V. Performance of a multiplexed dual analyte immunoassay for the early detection of non-small cell lung cancer. *J. Transl. Med.* **13**, 55 (2015).
- U.S. Preventive Services Task Force. Final recommendation statement: lung cancer: screening. *U.S. Preventive Services Task Force* <http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/lung-cancer-screening> (2013).
- Kauczor, H. U. *et al.* ESR/ERS white paper on lung cancer screening. *Eur. Respir. J.* **46**, 28–39 (2015).
- Wender, R. *et al.* American Cancer Society lung cancer screening guidelines. *C.A. Cancer J. Clin.* **63**, 107–117 (2013).
- Ruano-Ravina, A., Heleno, B. & Fernández-Villar, A. Lung cancer screening with low-dose CT (LDCT), or when a public health intervention is beyond the patient's benefit. *J. Epidemiol. Community Health.* **69**, 99–100 (2015).
- Ruchalski, K., Gutierrez, A., Genshaft, S., Abtin, F. & Suh, R. The evidence for low-dose CT screening of lung cancer. *Clin. Imaging.* **40**, 288–295 (2016).
- Sanz-Santos, J. *et al.* Usefulness of a lung cancer rapid diagnosis specialist clinic. Contribution of ultrasound bronchoscopy. *Arch. Bronconeumol.* **46**, 640–645 (2010).
- Huetó Pérez De Heredia, J. *et al.* Evaluation of the use of a rapid diagnostic consultation of lung cancer. Delay time of diagnosis and therapy. *Arch. Bronconeumol.* **48**, 267–273 (2012).
- Leiro-Fernández, V. *et al.* Effectiveness of a protocolized system to alert pulmonologists of lung cancer radiological suspicion. *Clin. Transl. Oncol.* **16**, 64–68 (2014).
- Detterbeck, F. C. *et al.* Invasive mediastinal staging of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest.* **132**, 202S–220S (2007).
- Silvestri, G. A. *et al.* Noninvasive staging of non-small cell lung cancer: ACCP evidenced-based clinical practice guidelines (2nd edition). *Chest.* **132**, 178S–201S (2007).
- Bosch, X., Aibar, J., Capell, S., Coca, A. & López-Soto, A. Quick diagnosis units: a potentially useful alternative to conventional hospitalisation. *Med. J. Aust.* **191**, 496–498 (2009).
- Gupta, S., Sukhal, S., Agarwal, R. & Das, K. Quick diagnosis units—an effective alternative to hospitalization for diagnostic workup: a systematic review. *J. Hosp. Med.* **9**, 54–59 (2014).
- Hensing, T. A. & Salgia, R. Molecular biomarkers for future screening of lung cancer. *J. Surg. Oncol.* **108**, 327–333 (2013).

18. I, H. & Cho, J. Y. Lung Cancer Biomarkers. *Adv. Clin. Chem.* **72**, 107–170 (2015).
19. Blanco-Prieto, S. *et al.* Serum calprotectin, CD26 and EGF to establish a panel for the diagnosis of lung cancer. *PLoS One.* **10**, e0127318; 10.1371/journal.pone.0127318 (2015).
20. Izbicka, E. *et al.* Plasma biomarkers distinguish non-small cell lung cancer from asthma and differ in men and women. *Cancer Genomics Proteomics.* **9**, 27–35 (2012).
21. Javidroozi, M., Zucker, S. & Chen, W. T. Plasma seprase and DPP4 levels as markers of disease and prognosis in cancer. *Dis. Markers.* **32**, 309–320 (2012).
22. Liu, P. J. *et al.* In-depth proteomic analysis of six types of exudative pleural effusions for nonsmall cell lung cancer biomarker discovery. *Mol. Cell Proteomics.* **14**, 917–932 (2015).
23. Sánchez-Otero, N. *et al.* Calprotectin: a novel biomarker for the diagnosis of pleural effusion. *Br. J. Cancer.* **107**, 1876–1882 (2012).
24. Yang, D. W. *et al.* Role of a serum-based biomarker panel in the early diagnosis of lung cancer for a cohort of high-risk patients. *Cancer.* **121**, 3113–3121 (2015).
25. Molina, R. *et al.* Assessment of a Combined Panel of Six Serum Tumor Markers for Lung Cancer. *Am. J. Respir. Crit. Care Med.* **193**, 427–437 (2016).
26. Chu, X. Y. *et al.* Diagnostic values of SCC, CEA, Cyfra21-1 and NSE for lung cancer in patients with suspicious pulmonary masses: a single center analysis. *Cancer Biol. Ther.* **11**, 995–1000 (2011).
27. Patz, E. F. Jr. *et al.* Biomarkers to help guide management of patients with pulmonary nodules. *Am. J. Respir. Crit. Care Med.* **188**, 461–465 (2013).
28. Jett, J. R. *et al.* Audit of the autoantibody test, EarlyCDT<sup>®</sup>-lung, in 1600 patients: an evaluation of its performance in routine clinical practice. *Lung Cancer.* **83**, 51–55 (2014).
29. Copas, J.B. Regression, prediction and shrinkage. *J. Roy. Statist. Soc. Series B.* **45**, 311–354 (1983).
30. Hastie, T., Tibshirani, R. & Friedman, J. Elements of Statistical Learning: Data Mining, Inference, and Prediction in *Springer Series in Statistics* (Springer, 2009).
31. Steyberg, E. W. Clinical Prediction Models: A practical approach to development, validation and updating in *Statistics for Biology and Health* (ed. Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., Wong, W.) (Springer, 2009).
32. Pesch, B. *et al.* Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case-control studies. *Int. J. Cancer.* **131**, 1210–1219 (2012).
33. Ikhtaire, S., Shajib, M. S., Reinisch, W. & Khan, W. I. Fecal calprotectin: its scope and utility in the management of inflammatory bowel disease. *J. Gastroenterol.* **51**, 434–446 (2016).
34. RStudio & Inc. shiny: *Web Application Framework for R* <http://cran.r-project.org/web/packages/shiny/> (2012).
35. Gosney, J. & Travis, W. D. Pathology and genetics: tumours of the lung, pleural, thymus and heart in *World Health Organization classification of tumours* (ed. Travis, W. D., Brambilla, E., Müller-Hermelink, H., Harris, C. C.) 76–77 (IARC Press, 2004).
36. Sánchez de Cos, J. *et al.* SEPAR guidelines for lung cancer staging. *Arch. Bronconeumol.* **47**, 454–465 (2011).
37. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc. B.* **57**, 289–300 (1995).
38. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B.* **58**, 267–288 (1996).
39. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
40. Akaike, H. Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Int. Symp. Information Theory Supp. to Problems of Control and Information Theory.* 267–281 (1972).
41. Schwarz, G. E. Estimating the dimension of a model. *Annals of Statistics.* **6**, 461–464 (1978).
42. Roskoski, R. Jr. The ErbB/HER family of protein-tyrosine kinases and cancer. *Pharmacol. Res.* **79**, 34–74 (2014).
43. De Meester, I., Korom, S., Van Damme, J. & Scharpé, S. CD26, let it cut or cut it down. *Immunol. Today.* **20**, 367–375 (1999).
44. Wesley, U. V., Tiwari, S. & Houghton, A. N. Role for dipeptidyl peptidase IV in tumor suppression of human non small cell lung carcinoma cells. *Int. J. Cancer.* **109**, 855–866 (2004).
45. Ghavami, S. *et al.* S100A8/A9: a Janus-faced molecule in cancer therapy and tumorigenesis. *Eur. J. Pharmacol.* **625**, 73–83 (2009).
46. Hiratsuka, S., Watanabe, A., Aburatani, H. & Maru, Y. Tumour-mediated upregulation of chemoattractants and recruitment of myeloid cells predetermines lung metastasis. *Nat. Cell Biol.* **8**, 1369–1375 (2006).
47. Kessenbrock, K., Plaks, V. & Werb, Z. Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell.* **141**, 52–67 (2010).
48. Li, M. *et al.* Prognostic significance of matrix metalloproteinase-1 levels in peripheral plasma and tumour tissues of lung cancer patients. *Lung Cancer.* **69**, 341–347 (2010).
49. Ulivi, P. *et al.* MMP-7 and fcDNA serum levels in early NSCLC and idiopathic interstitial pneumonia: preliminary study. *Int. J. Mol. Sci.* **14**, 24097–24112 (2013).
50. Zhang, Y. *et al.* Detection of circulating vascular endothelial growth factor and matrix metalloproteinase-9 in non-small cell lung cancer using Luminex multiplex technology. *Oncol. Lett.* **7**, 499–506 (2014).
51. Jumper, C., Cobos, E. & Lox, C. Determination of the serum matrix metalloproteinase-9 (MMP-9) and tissue inhibitor of matrix metalloproteinase-1 (TIMP-1) in patients with either advanced small-cell lung cancer or non-small-cell lung cancer prior to treatment. *Respir. Med.* **98**, 173–177 (2004).
52. Hammarström, S. The carcinoembryonic antigen (CEA) family: structures, suggested functions and expression in normal and malignant tissues. *Semin. Cancer Biol.* **9**, 67–81 (1999).
53. Moll, R., Franke, W. W., Schiller, D. L., Geiger, B. & Krepler, R. The catalog of human cytokeratins: patterns of expression in normal epithelia, tumors and cultured cells. *Cell.* **31**, 11–24 (1982).

## Acknowledgements

This study was supported by Instituto de Salud Carlos III (project PS09-00405), and Xunta de Galicia (INBIOMED 2012-273 and GRC2014/019, FEDER support included). S. Blanco-Prieto was supported by Spanish Ministry of Science and Innovation (fellowship FPU). M. Rodríguez-Girondo was supported by the Spanish Ministry of Science and Innovation (Grant MTM2011-23204) and Grant MIMomics of the European Union's Seventh Framework Programme (FP7-Health-F5-2012) number 305280.

Authors would like to thank L. Barcia-Castro, from Department of Biochemistry, Genetics and Immunology, Universidad de Vigo (Spain) for technical assistance. The samples used in this work belong to the Biobank from EOXI Vigo (RETIC-FISISCIIRD09/0076/00011).

## Author Contributions

F.J.R.B., M.P., A.F.V. designed the study; S.B.P., L.D., L.V.I., M.I.B.R. acquired the data; S.B.P., L.D., M.R.G., M.P. analysed and interpreted the data; S.B.P., L.D., M.R.G., L.V.I., M.P. prepared the manuscript; all authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Blanco-Prieto, S. *et al.* Highly Sensitive Marker Panel for Guidance in Lung Cancer Rapid Diagnostic Units. *Sci. Rep.* 7, 41151; doi: 10.1038/srep41151 (2017).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017