Expert Systems with Applications 39 (2012) 10851–10856

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Development of an optimized multi-biomarker panel for the detection of lung cancer based on principal component analysis and artificial neural network modeling

José Miguel Flores-Fernández [a], Enrique J. Herrera-López [a], Francisco Sánchez-Llamas [b], Antonio Rojas-Calvillo [c], Paula Anel Cabrera-Galeana [c], Gisela Leal-Pacheco [a] María Guadalupe González-Palomar [a], R. Femat [d], Moisés Martínez-Velázquez [a,]*

[a] *Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco, AC, Av. Normalistas 800, C.P. 44270 Guadalajara, Jalisco, Mexico*
[b] *OPD Antiguo Hospital Civil de Guadalajara "Fray Antonio Alcalde", Servicio de Fisiología Pulmonar e Inhaloterapia, Hospital 278, C.P. 44280 Guadalajara, Jalisco, Mexico*
[c] *Coordinación de Oncología Médica, Centro Oncológico Estatal ISSEMYM, Avenida Solidaridad las Torres 101, C.P. 50180 Toluca, Estado de Mexico, Mexico*
[d] *División de Matemáticas Aplicadas, IPICyT, Camino a la presa San José 2055, Lomas 4a Secc, C.P. 78216 San Luis Potosí, Mexico*

ARTICLE INFO

Keywords:
Biomarkers
Artificial neural network
Principal component analysis
Diagnosis
Lung cancer

ABSTRACT

Lung cancer is a public health priority worldwide due to the high mortality rate and the costs involved. Early detection of lung cancer is important for increasing the survival rate, however, frequently its diagnosis is not made opportunely, since detection methods are not sufficiently sensitive and specific. In recent years serum biomarkers have been proposed as a method that might enhance diagnostic capabilities and complement imaging studies. However, when used alone they show low sensitivity and specificity because lung cancer is a heterogeneous disease. Recent reports have shown that simultaneous analysis of biomarkers has the potential to separate lung cancer patients from control subjects. However, it has become clear that a universal biomarker panel does not exist, and optimized panels need to be developed and validated in each population before their application in a clinical setting. In this study, we selected 14 biomarkers from literature, whose diagnostic or prognostic value had been previously demonstrated for lung cancer, and evaluated them in sera from 63 patients with lung cancer and 87 non-cancer controls (58 Chronic Obstructive Pulmonary Disease (COPD) patients and 29 current smokers). Principal component analysis and artificial neural network modeling allowed us to find a reduced biomarker panel composed of Cyfra 21.1, CEA, CA125 and CRP. This panel was able to correctly classify 135 out of 150 subjects, showing a correct classification rate for lung cancer patients of 88.9%, 93.3% and 90% in training, validation and testing phases, respectively. Thus, sensitivity was increased 18.31% (sensitivity 94.5% at specificity 80%) with respect to the best single marker Cyfra 21.1. This optimized panel represents a potential tool for assisting lung cancer diagnosis, therefore it merits further consideration.

1. Introduction

Lung cancer is a public health priority worldwide due to the high mortality rate and the costs involved (Hernández-García, Sáenz-González, & González-Celador, 2010). In 2004, the lung cancer mortality rate in Mexico was approximately 6.49 per 100,000 inhabitants and it has been estimated that about 10,000 people died by this disease in 2010 (Franco-Marina & Villalba-Caloca, 2001; Ruíz-Godoy et al., 2007). Early detection of lung cancer is important for increasing the survival rate, and diverse methods have been used for detecting the disease, such as chest radiography, low-dose computed tomography, sputum cytology and bronchoscopy, among others (Castleberry, Smith, Anderson, Rotter, & Grannis, 2009; Ghosal, Kloer, & Lewis, 2009). However, frequently lung cancer diagnosis is not made opportunely, since detection methods are not sufficiently sensitive and specific (Bajtarevic et al., 2009).

Serum biomarkers have been proposed as a complementary method for cancer detection (Fernández et al., 2007). Proteins such as cytokeratin 19 fragment (CYFRA 21-1), carcinoembryonic

* Corresponding author. Address: CIATEJ, AC. Av. Normalistas 800, Col. Colinas de la Normal, C.P. 44270, Guadalajara, Jalisco, Mexico. Tel.: +52 3333455200x1370; fax: +52 3333455200x1001.

E-mail addresses: moisesmartinezv@yahoo.com.mx, mmartinez@ciatej.net.mx (M. Martínez-Velázquez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.008

antigen (CEA), squamous cell carcinoma antigen (SCCA), tissue polypeptide antigen (TPA), cancer antigen 125 (CA-125), squamous cell carcinoma (SCC), progastrin-releasing peptide (ProGRP), neuron-specific enolase (NSE), among others, have shown to be potentially useful for diagnosis of lung cancer (Cho, 2007). Nevertheless, they have not provided sufficient specificity and sensitivity when used alone (Bajtarevic et al., 2009). Combining different biomarkers might improve the diagnostic power (Chu et al., 2011; Farlow et al., 2010a, 2010b; Leidinger et al., 2010; Patz et al., 2007), however, identifying an optimal set of biomarkers, which may significantly impact lung cancer detection, is a difficult task.

In recent years, artificial neural networks (ANNs), fuzzy logic and genetic algorithms have been proposed as auxiliary tools in medicine. Fuzzy logic could increase the sensitivity and specificity of biomarkers for lung cancer (Schneider, Bitterlich, Kotschy-Lang, Raab, & Hans-Joachim, 2007; Schneider et al., 2002; Schneider et al., 2003; Zhang, Zhou, Liu, & Harrington, 2006), whereas ANNs may play an important role in lung cancer, in morphologically differentiating malignant from benign cells (Zhou, Jiang, Yang, & Chen, 2002) and in detection of pulmonary nodules from computed tomography chest images (Gomathi & Thangaraj, 2010). More recently, ANNs were used to discriminate between clinical stage I and IV in melanoma (Lancashire et al., 2005), and to discriminate lung cancer from benign lung disease using ANN with six serum tumor markers and 19 parameters of basic information from patients (Feng, Wu, Wu, Nie, & Ni, 2011; Wu et al., 2011).

In this study, we selected an assortment of biomarkers from literature, which previously demonstrated diagnostic or prognostic value for lung cancer, and applied a combination of multivariate analysis tools with ANNs modeling in order to find an optimal and minimum panel of biomarkers to discriminate lung cancer patients from high-risk people .

2. Materials and methods

2.1. Patient populations

150 subjects were enrolled from Antiguo Hospital Civil de Guadalajara and Centro Oncológico Estatal ISSEMyM, and divided in two groups. The first group included 63 consecutive lung cancer patients (37 men, 26 women) with newly diagnosed lung cancer. The second group was composed of patients at high risk for developing lung cancer but with no history of lung disease – 58 Chronic Obstructive Pulmonary Disease (COPD) patients (40 men, 18 women) and 29 current smokers (17 men, 12 women) with no history of lung disease. Approval was obtained from the corresponding ethics committee and all participants gave written informed consent. Histological diagnosis of primary lung cancer was established according to the revised classification of lung tumors of the World Health Organization and the International Association for Lung Cancer Study (Travis, Brambilla, Mueller-Hermelink, & Harris, 2004). In accordance with current Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines, COPD was defined by a post-bronchodilator FEV1/FVC ratio < 0.70. Samples and health information were labeled using unique identifiers to protect subject confidentiality.

2.2. Measurement of serum biomarker concentrations

Ten ml of blood from each patient were collected in serum separator tubes, processed immediately and separated by centrifugation at 3,000 rpm at room temperature for 10 min. The separated serum was then aliquoted and stored at −70 °C for future analysis. Fourteen biomarkers which had previously been associated with lung cancer were chosen. Levels of Matrix metalloproteinase-1 (MMP1), Matrix metalloproteinase-9 (MMP9; R&D Systems, Minneapolis, MN), Urokinase plasminogen activator (uPA), Transferrin (TF), α1-Antitrypsin (AT), Haptoglobulin (HPT), Apolipoprotein A-I (APOAI), Retinol-binding protein (RBP), C-reactive protein (CRP; ASSAYPRO, St. Charles, MO), Cancer antigen 125 (CA125), Carcinoembryonic antigen (CEA), Neuron-specific enolase (NSE; ALPCO Diagnostics, Salem, NH), Cytokeratin 19 fragment (Cyfra 21-1; DRG Instruments GmbH, Germany), and YKL40 (Quidel Corporation, San Diego, CA) were measured in serum samples with a solid phase sandwich enzyme-linked immunosorbent assay (ELISA), using commercially available human ELISA assays and in accordance with the kits' directions. The absorbance specified in the protocols was measured by an automatic ELISA reader (Bio-Rad Laboratories, Philadelphia, PA). Results were converted from the mean absorbance of duplicate wells after subtraction of background values. Recombinant human MMP1, MMP9, CA125, CEA, NSE, Cyfra 21-1, YKL40, uPA, HPT, TF, CRP, APO-AI, RBP and AT proteins were used as standards. The standard curve was prepared simultaneously with the test samples.

2.3. Statistical analysis

Differences between groups were calculated by means of a non-parametric test (Mann–Whitney's U-test). Values of $p < 0.05$ were considered significant. Statistical analyses were performed using SigmaStat 8.0.

2.4. Artificial neural network (ANN) modeling

ANNs are tools of artificial intelligence intended to imitate the complex operation of organizing and processing information of the neurons in the brain. ANNs can identify patterns that correlate strongly a set of data which correspond to a class by a learning process, in which interneuron connection weights are used to store knowledge about specific features identified within the data (Lancashire et al., 2005). A common ANN is the Multilayered Perceptron (MLP) which is composed of three layers as shown in Fig. 1. The ANN is trained entering information from the input layer through the hidden and output layers of the network. The hidden and output layers are trasformed by a validation function $f(\theta)$ that could be the heaviside step, linear, sigmoidal, hyperbolic tangent, among others. The value of the output is compared with a known target vector and the difference is computed as an error. If the output is different than the target vector then the connection weights (W_{ij}, W_{ik}) are readjusted backwards by means of a backpropagation training algorithm which minimizes the error. The procedure is repeated until the output of the ANN reaches a desired error value. It is not the aim of an ANN to perfectly represent the data used during the training set but to extract relevant information which allows modeling different conditions not used during the training process. The functionality of the trained ANN is validated by passing data not used during the training. The

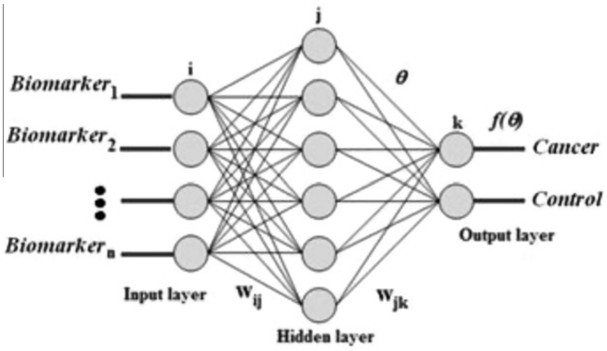

Fig. 1. Structure for the Multilayer Perceptron network used to classify lung cancer patients and controls with biomarker as input of the ANN.

software used for training the ANN was Matlab 7.9 and the Neural Networks Toolbox 7.9.

Preprocessing: The 14 marker values from all subjects were preprocessed before being used as inputs to the ANN. The funcion "mapstd" was used to normalize marker values to a mean value 0 with a standard deviation of 1.

Training: From previous results where different ANNs were tested (data not shown), it was determined that the pattern recognition network (PRN) might be used for classifying lung cancer and control patients (Flores et al., 2011). A PRN is a feedforward network trained to classify inputs according to target classes. In order to determine the best architecture for the PRN, structures with one hidden layer (with 3, 5, 8, 10, 12, 15 and 20 nodes) and two hidden layers (with 3–2, 5–3, 2–6, 6–3 and 5–5 nodes) were tested. Two neurons were set in the output layer, representing cancer patients and control subjects, respectively. The optimal architecture was chosen as the one resulting with the lowest training error. The nonlinear tangential sigmoidal activation function for the hidden neurons was "tansig" while the linear activation function for the output neurons was "purelin" in the range 0 to 1. The PRN was trained with the algorithm "trainrp". The stopping criterion was set as the mean squared error (MSE) less than 0.09, where

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(t_i - \theta_i)^2 \tag{1}$$

and t is the target vector, θ is the output of the ANN and N is the amount of data.

The whole data set of biomarkers information was randomly divided as follows: 60% for training and 20% for validation; this matrix was used in all the experiments. In addition, 20% of samples not used during training the ANN were used for testing. To avoid over-fitting a 10-fold cross-validation was used.

2.5. Principal component analysis (PCA)

In order to determine uncorrelated biomarkers that better explain the variability observed in the data, a PCA was applied to biomarkers that showed significant differences between groups, by using Simca-P + 12.0^{TM}. The best biomarkers were used to train the ANN.

2.6. Detection capability comparison

The sensitivity of the ANN with the best combination of biomarkers was compared with the best single biomarker by means of Receiver Operating Characteristic (ROC) curves at a specificity of relevant clinical value (80%). The ROC curves were graphed using SigmaPlot^{TM} 10.0. A p-value < 0.05 (Kolmogorov-Smirnof test) was considered significant.

2.7. Discriminant analysis

To compare the performance of the ANN against multivariate statistics,we used a classical parametric statistical method (linear discriminant analysis, LDA). LDA provides certain linear combinations of the original variables called discriminant functions, which are likely to improve on the discrimination between the groups than any individual variable. LDA was performed using Statgraphics Centurion XV©, which includes stepwise forward and backward, in attempt to simplify the classification model, obtaining the same improvement with both algorithms. The criteria for adding or removing variables from the model was statistical value *F*. Even though we tried to convert the experimental data, most of them did not present normal distribution, due to the fact that protein concentrations showed large differences within groups. Even standard deviations of the data showed no homogeneity between groups.

3. Results

3.1. Analysis of individual serum biomarkers

As a first step in the effort to establish a novel multi-analyte serum test for lung cancer detection, we selected an assortment of 14 biomarkers and determined their concentrations in serum from 63 lung cancer patients and 87 control subjects. Table 1 summarizes demographic and clinical characteristics of patients. These groups are not significantly different regarding age and smoking index. Lung cancer patients showed higher serum concentrations of MMP1, MMP9, AT, HPT, CA125, CEA, Cyfra 21-1, NSE, CRP and YKL-40 than controls, whereas concentrations of TF, APOAI and RBP were lower in lung cancer patients, and uPA showed no difference (Table 2).

3.2. Artificial neural network modeling

3.2.1. ANN trained with fourteen biomarkers

A multilayer pattern recognition ANN that we named ANN$_1$ was trained, validated and tested with 90, 30 and 30 samples, respectively. Each sample contained the Normalized concentration of the whole set of 14 biomarkers (MMP1, MMP9, uPA, TF, AT, HPT, CA125, CEA, Cyfra 21-1, NSE, APOAI, RBP, CRP and YKL40), which were used as inputs for the ANN. It was determined that 10 neurons in the hidden layer and two neurons in the output layer gave the best performance for the ANN$_1$. All trainings were made in triplicate to assure that the best architecture was chosen. A MSE of 0.063 was reached with the optimum structure. Using ANN$_{1,}$ 133 of 150 subjects were classified correctly (including lung cancer patients and control subjects), which represents a coefficient of determination R^2 = 92.2% (83/90), 93.3% (28/30) and 73.3% (22/30) for training, validation and testing phases, respectively (Table 3).

3.2.2. Reduction in biomarkers number (inputs) and evaluation of the ANN performance

Although the performance of ANN$_1$ with 14 biomarkers was acceptable, we explored whether the reduction in the number of biomarkers would affect the ANN performance. Mann–Whitney's U-test was carried out for all the biomarkers concentrations, determining significant differences (p < 0.05) in thirteen biomarkers (10 increased and 3 decreased in lung cancer group, Table 2). Then, a prinicpal components analysis (PCA) was carried out over these

Table 1
Demographic and clinical profiles of lung cancer patients and controls.

Demographic	Controls (n = 87)	Lung cancer (n = 63)	*p*-value
Age (years)	61 ± 15.2	64 ± 13.9	0.267
Range (age)	19–92	31–89	–
Female	30	26	–
Male	57	37	–
Active smoking	65	50	–
Passive smoking	0	2	–
Cooking with wood	2	6	–
Smoking/Wood	9	2	–
Unknown cause		3	–
Smoking index	30 ± 27.6	36 ± 30.4	0.412
NSCLC	–	50	–
SCLC	–	9	–
N/A	–	4	–
III	–	8	–
IV	–	23	–

N/A: Not available

Table 2
Serum concentration of 14 biomarkers evaluated in lung cancer patients and control subjects.

Protein	Control 25–75%		Lung cancer 25–75%		*p*-value
	Median	Quartile	Median	Quartile	
MMP1 (ng/ml)	17.48	6.56–23.61	26.48*	18.18–34.1	<0.0001
MMP9 (ng/ml)	551.81	369.46–747.24	900.38*	470.02–1299.27	<0.0001
uPA (ng/ml)	1.49	0.7–4.71	1.8	0.33–5.12	0.7159
TF (μg/ml)	5123.76*	1639.0–11371.4	1585	817.2–6307.86	<0.0001
AT (μg/ml)	1211.61	621.19–1691.84	2091.36*	913.06–2571.92	<0.0001
HPT (μg/ml)	2362.71	1230.2–3392	4722.2*	2887.2–7581.4	<0.0001
CA125 (U/ml)	4.71	1.5–11–43	25.31*	14.82–72.5	0.0000
CEA (ng/ml)	1.52	0.71–2.23	2.82*	1.13–18.93	0.0002
Cyfra 21-1 (ng/ml)	0.79	0.79–1.12	4.2*	1.28–9.48	0.0000
NSE (ng/ml)	10.36	5.84–19.59	12.69*	8.71–23.63	0.0202
APOAI (μg/ml)	8671.1*	3450.06–11865.2	6254.24	1847.53–10395.8	0.0293
RBP (μg/ml)	36.26*	25.67–314.48	24.43	21.34–299.2	0.0371
CRP (μg/ml)	7150	2813.3–16532.8	24833.9*	19350–32000.00	<0.0001
YKL40 (μg/ml)	171.88	110.21–326.23	353.79*	238.51–591.74	<0.0001

* Significantly different by Mann–Whitney U test.

Table 3
Correct classification rate of the developed ANNs.

	ANN_1		ANN_2		ANN_3	
	%	n	%	n	%	n
Train	92.2	83	86.7	78	88.9	80
Val	93.3	28	96.7	29	93.3	28
Test	73.3	22	86.7	26	90	27
Total	88.7	133	88.7	133	90	135

biomarkers (the PCA is a multivariate, non-parametric method for extracting relevant information from confusing data sets. The idea behind the PCA was to determine those biomarkers that best represent the majority of lung cancer cases). Fig. 2 shows the weight of biomarkers regarding to their principal components. Ten biomarkers have influence on the lung cancer group, however, the subgroup CEA, CA125, MMP9, Cyfra 21-1, YKL40 and CRP clearly have greater influence, since these biomarkers have the largest weights on the first and second components. PCA results show that when these proteins have higher concentration, they tend to belong to the lung cancer group. The plot of the PCA shows a marked difference on the lung cancer patients and control subjects distribution (Fig. 3).

After the initial vector of 14 biomarkers was reduced to six, a new net named ANN_2 was trained with these proteins. A multilayered pattern recognition ANN with 5 neurons in the hidden layer was chosen as the best configuration in classifying patients. The vector previously used for training, validation and testing was applied to the ANN_2. This network correctly classified 133 of 150 subjects, which represents a determination coefficient of 86.7% (78/90), 96.7% (29/30) and 86.7% (26/30) for training, validation and testing phases, respectively (Table 3). In this case, decreasing the number of biomarkers did not affect the network performance.

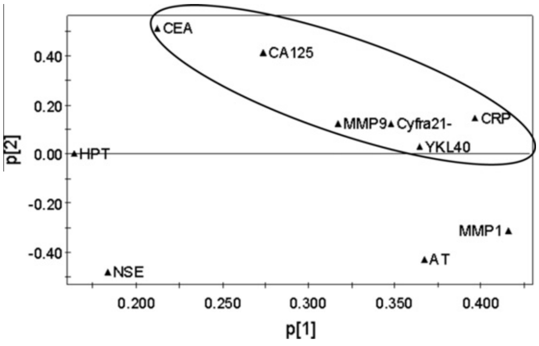

Fig. 2. Principal component analysis performed with biomarkers that showed higher serum concentrations for lung cancer group.

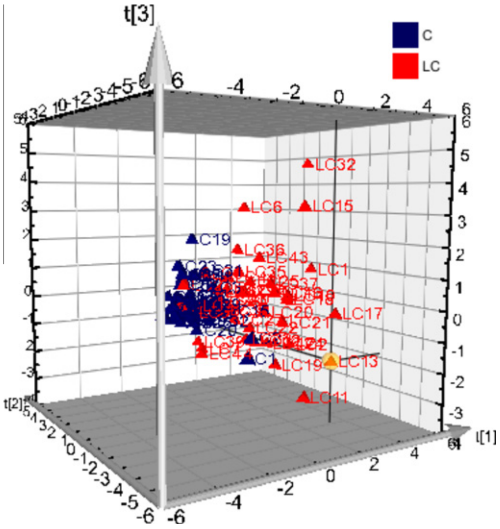

Fig. 3. Distribution of the lung cancer group (LC) and high-risk population (C) with regard to its main components.

3.2.3. Development of an optimized biomarker panel based on ANN modeling

Finally, in order to find the most reduced panel, all the 56 possible combinations with the six biomarkers derived from the PCA were used to create different ANNs. From the ANNs tested in this stage, it was determined that the new net named ANN_3 with five neurons in the hidden layer and the biomarkers combination Cyfra 21-1, CEA, CA125 and CRP had the best performance. ANN_3, trained with 4 biomarkers, correctly classified 135 out of 150 subjects. Thus, ANN_3 performed similarly to ANN_1 and ANN_2. The classification rate was slightly improved since two additional cancer patients were classified correctly. The correct classification rate for patients was 88.9% (80/90), 93.3% (28/30) and 90% (27/30) for training, validation and testing stages, respectively (Table 3).

The detection capability of each single biomarker and the performance of the ANNs were evaluated by ROC curves. Cyfra

21-1 was the best single protein in classifying lung cancer with a sensitivity of 76.19% at a specificity of 80%. The ROC curve and the sensitivity at a specificity of 80% for Cyfra 21-1 were compared with the ROC curves resulting from the overall performance from ANN_1, ANN_2 and ANN_3 (Fig. 4). ANN_1 and ANN_2, trained with 14 and 6 markers, respectively, increased the sensitivity by 19.71% (sensitivity 95.9% at specificity 80%) and 13.21% (sensitivity 89.4% at specificity 80%), respectively, whereas ANN_3 increased the sensitivity 18.31% (sensitivity 94.5% at specificity 80%) with respect to the best single marker Cyfra 21-1. The area under the curve values for Cyfra 21-1, ANN_1, ANN_2 and ANN_3 were 0.85, 0.934, 0.933 and 0.939, respectively (Fig. 4).

3.2.4. Comparison of the ANN with a statistical method

In order to compare the classification power of the ANNs against a statistical tool, a discriminant analysis (DA) was carried out. The DA ($p<0.0001$) correctly classified 76.2% of patients from the lung cancer group and 93.1% subjects from the control group, using the biomarkers AT, MMP-9, HPT, Cyfra 21-1, YKL40, CEA, RBP and TF (Fig. 4). ANN_3 had a superior overall classification rate for lung cancer patients (82.54%) and for high-risk control subjects (96.55%).

4. Discussion

Lung cancer causes more deaths than any other cancer. Despite scientific advances in the lung cancer diagnosis, treatment and prognosis, mortality has not changed for several decades (Patz et al., 2007). Because the disease is asymptomatic in early stages (Lu et al., 2004), over 75% cases are diagnosed when the disease is in an advanced stage, where treatment options are limited and, as a consequence, the survival rate at 5 years is less than 15% (Farlow et al., 2010a). Therefore, detecting lung cancer in early stages, when it is possible to resect the tumor, would achieve healing and thereby reduce mortality.

Computed tomography (CT) has been proposed as a screening method for early detection in high-risk populations, due to its high sensitivity in detecting pulmonary nodules (The International Early Lung Cancer Action Program Investigators, 2006). Recently released guidelines from the National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology. Lung Cancer Screening. Version 1.2012. Retrieved December (2011) in the United States, strongly recommend the use of low-dose computed tomography screening for select individuals at high risk for the disease. However, CT has a low specificity in that only a small percentage of nodule-positive patients will develop lung cancer, and furthermore, repeated radiation may promote carcinogenesis (Brenner & Hall, 2007; Humphrey, Teutsch, & Johnson, 2004; Swensen et al., 2005).

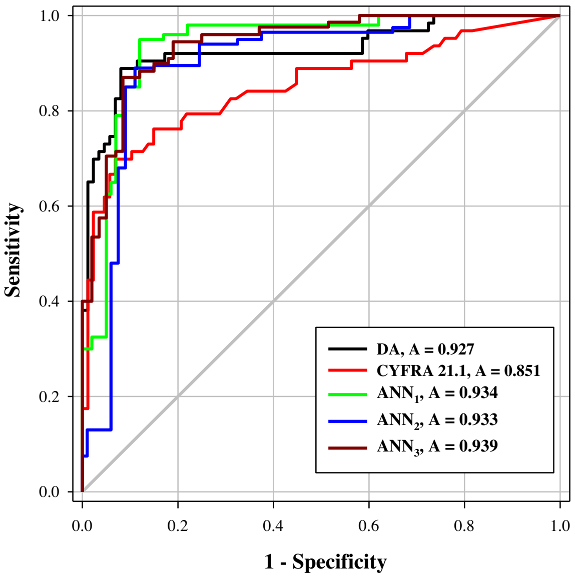

Fig. 4. Comparison of the performance (sensitivity) from combined biomarkers by ANN, the best individual biomarker, Cyfra 21-1, and discriminant analysis (DA). ANN_1 trained with 14 biomarkers; ANN_2 trained with 6 biomarkers; ANN_3 trained with 4 biomarkers.

Biological markers that define high-risk patients would enhance diagnostic capabilities and complement imaging studies because they can easily be detected in biological fluids using minimally invasive procedures. However, when used alone they show low sensitivity and specificity because lung cancer is a heterogeneous disease (He et al., 2007; Tureci et al., 2006). Recent reports have shown that simultaneous analysis of biomarkers has the potential to separate lung cancer patients from control subjects (Farlow et al., 2010a; Farlow et al., 2010b; Lee et al., 2011; Leidinger et al., 2010; Patz et al., 2007; Schneider et al., 2002; Zhong et al., 2006). However, it has become clear that a universal biomarker panel does not exist, and optimized panels need to be developed and validated in each population before their application in a clinical setting. For that reason, in the present study we selected an assortment of biomarkers from literature, evaluated their concentrations and applied a combination of PCA with ANN modeling to determine the biomarker combination that best discriminates lung cancer patients from control subjects.

ANN_1 trained with fourteen biomarkers correctly classified 133 of 150 samples. However, this panel of 14 markers may not be the best option for lung cancer diagnosis, due to the high costs of the test. Therefore, it would be desirable to reduce the number of biomarkers without detriment to the ANN performance. Moreover, as a part of the optimization process, it is important to discard non-informative biomarkers that could perturb the classifier. Applying PCA, the number of biomarkers was reduced from 14 to just 6. ANN_2 did not lose accuracy since the eight biomarkers eliminated were redundant in information. Recently, Wu et al. (2011) performed a study in Chinese people, training a Back Propagation network (6, 3 and 1 neurons in the input, hidden and output layer, respectively) with 50 lung cancer patients, 40 benign lung disease patients and 50 normal people. Their ANN had a satisfactory performance with 6 biomarkers (β_2-microglobulin, CEA, gastrin, CA125, NSE, soluble interleukin-6 receptor) and three metal ions (Cu^{2+}, Zn^{2+} and Ca^{2+}), achieving a prediction rate of 85% in the test phase. They used several biomarker numbers (3, 4, 6 and 12) as input for the ANN, and similarly, they concluded that increasing biomarker number to more than 6 did not significantly increase the prediction rate. In a related study, Feng et al. (2011) employed an ANN with the same 6 biomarkers described by Wu et al. (2011), adding 19 parameters with information such as risk factors, symptoms, smoking, chemical exposure, kitchen environment, and so on. This ANN had 25 neurons in the input layer and 15 in the hidden layer. Their ANN reached a prediction rate of 87.3% in the test phase, similar to our results.

Taking advantage of the ANN's capabilities to find the most reduced panel, all the possible combinations with the six biomarkers derived from the PCA were used to create 56 different ANNs. From these, ANN_3 with five neurons in the hidden layer and the biomarker combination Cyfra 21-1, CEA, CA125 and CRP presented the best performance. It was found that decreasing the biomarkers to four did not affect the network performance, in fact, two additional patients in the cancer group were correctly classified, perhaps because the other two markers gave redundant information during ANN training. ANN_3 had the best sensitivity at a specificity of 80%, compared with the other nets and the best single marker (see Fig. 4). On the other hand, ANN_3 required fewer markers than needed by discriminant analysis to separate the study groups (four and eight proteins, respectively), which would represent a reduction in costs associated. The robustness of ANN_3 might be improved by adding samples from new subjects, while keeping the normalizing criteria to train the network. In addition, clinical information such as age, gender, smoking, nodules, etcetera, could be included to improve the ANN performance.

In summary, we presented a strategy based on ANN modeling to find the best biomarker combination to distinguish lung cancer patients from high-risk people. This optimized panel is, however, perfectible and awaits additional markers that enhance its diagnostic accuracy, as well as further consideration for future studies.

Acknowledgments

This work was supported by grant CONACYT-FOSISS-2008-1-87628 from Consejo Nacional de Ciencia y Tecnología to MMV. JMFF, GLP and MGGP received a MSc scholarship from CONACYT. We thank Kirk Allen of Peace Corps Mexico for revising the language of manuscript and Karen Isabel Flores Fernandez for helping in blood specimen collection and processing.

References

Bajtarevic, A., Ager, C., Pienz, M., Klieber, M., Schwarz, K., Ligor, M., et al. (2009). Noninvasive detection of lung cancer by analysis of exhaled breath. *BMC Cancer, 9*(348), 1–16.

Brenner, D. J., & Hall, E. J. (2007). Computed tomography-an increasing source of radiation exposure. *The New England Journal of Medicine, 357*, 2277–2284.

Castleberry, A. W., Smith, D., Anderson, C., Rotter, A. J., & Grannis, F. W. Jr., (2009). Cost of a 5-year lung cancer survivor: symptomatic tumour identification vs proactive computed tomography screening. *British Journal of Cancer, 101*, 882–896.

Cho, W. (2007). Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. *Biomedicine & Pharmacotherapy, 61*, 515–519.

Chu, X. Y., Hou, X. B., Song, W. A., Xue, Z. Q., Wang, B., & Zhang, L. B. (2011). Diagnostic values of SCC, CEA, Cyfra21-1 and NSE for lung cancer in patients with suspicious pulmonary masses: a single center analysis. *Cancer Biology & Therapy, 11*, 995–1000.

Farlow, E. C., Patel, K., Basu, S., Lee, B. S., Kim, A. W., Coon, J. S., et al. (2010a). Development of a multiplexed tumor-associated autoantibody-based blood test for the detection of non-small cell lung cancer. *Clinical Cancer Research, 16*, 3452–3462.

Farlow, E. C., Vercillo, M. S., Coon, J. S., Basu, S., Kim, A. W., Faber, L. P., et al. (2010b). A multi-analyte serum test for the detection of non-small cell lung cancer. *British Journal of Cancer, 103*, 1221–1228.

Feng, F., Wu, Y., Wu, Y., Nie, G., & Ni, R. (2011). The effect of artificial neural networks model combined with six tumor markers in auxiliary diagnosis of lung cancer. *Journal of Medical Systems*. http://dx.doi.org/10.1007/s10916-011-9775-1.

Fernández, S. A., Martínez, P. A., Gaspar, M. J., Filella, X., Molina, R., & Ballesta, A. M. (2007). Marcadores tumorales serológicos. *Química Clínica, 26*, 77–85.

Flores, J. M., Herrera, E., Leal, G., González, M. G., Sánchez, F., Rojas A., Cabrera, P. A., Femat, R., Martínez-Velázquez, M. (2011). Artificial neural network-based serum biomarkers analysis improves sensitivity in the diagnosis of lung cancer. In: *Vol. 33 V Latin American Congress on Biomedical Engineering CLAIB 2011, IFMBE Proceedings 33*. La Habana, Cuba.

Franco-Marina, F., & Villalba-Caloca, J. (2001). La epidemia de cáncer pulmonar en México. *Revista del Instituto Nacional de Enfermedades Respiratorias, 14*, 207–214.

Ghosal, R., Kloer, P., & Lewis, K. E. (2009). A review of novel biological tools used in screening for the early detection of lung cancer. *Postgraduate Medical Journal, 85*, 358–363.

Gomathi, M., & Thangaraj, P. (2010). A computer aided diagnosis system for lung cancer detection using machine learning technique. *European Journal of Scientific Research, 2*, 5770–5779.

He, P., Naka, T., Serada, S., Fujimoto, M., Tanaka, T., Hashimoto, S., et al. (2007). Proteomics-based identification of alpha-enolase as a tumor antigen in non-small lung cancer. *Cancer Science, 98*(8), 1234–1240.

Hernández-García, I., Sáenz-González, M., & González-Celador, R. (2010). Mortality attributable to smoking in Spain in 2006. *Anales del Sistema Sanitario de Navarra, 33*, 23–33.

Humphrey, L. L., Teutsch, S., & Johnson, M. (2004). Lung cancer screening with sputum cytologic examination, chest radiography, and computed tomography: an update for the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 140*(9), 740–753.

Lancashire, L., Ugurel, S., Creaser, C., Schadendorf, D., Rees, R., Ball, G. (2005). Utilizing artificial neural networks to elucidate serum biomarker patterns which discriminate between clinical stages in melanoma. In: *2nd IEEE Symposium on Computational Intelligence in Bioformatics and Computational Biology* (pp. 455–460). California, USA: La Jolla.

Lee, H. J., Kim, Y., Park, T., Shin, P. J., Kang, Y. S., & Kim, K. N. (2011). A novel detection method of non–small cell lung cancer using multiplexed bead-based serum biomarker profiling. *The Journal of Thoracic and Cardiovascular Surgery, 12*. http://dx.doi.org/10.1016/j.jtcvs.2011.10.046.

Leidinger, P., Keller, A., Heisel, S., Ludwig, N., Rheinheimer, S., Klein, V., et al. (2010). Identification of lung cancer with high sensitivity and specificity by blood testing. *Respiratory Research, 11*, 18.

Lu, C., Soria, J. C., Tang, X., Xu, X. C., Wang, L., Mao, L., et al. (2004). Prognostic factors in resected stage I non-small-cell lung cancer: a multivariate analysis of six molecular markers. *Journal of Clinical Oncology, 22*(22), 4575–4583.

National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology. Lung Cancer Screening. Version 1.2012. Retrieved December 2011 from http://www.nccn.org/professionals/physician_gls/f_guidelines.asp.

Patz, E. F., Jr., Campa, M. J., Gottlin, E. B., Kusmartseva, I., Xiang, R. G., & Herndon, J. E. (2007). Panel of serum biomarkers for the diagnosis of lung cancer. *Journal of Clinical Oncology, 25*, 5578–5583.

Ruíz-Godoy, L., Rizo, R. P., Sánchez, C. F., Osornio-Vargas, A., García-Cuellar, C., & Meneses, G. A. (2007). Mortality due to lung cancer in Mexico. *Lung cancer, 58*, 184–190.

Schneider, J., Bitterlich, N., Velcovsky, H. G., Morr, H., Katz, N., & Eigenbrodt, E. (2002). Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer. *International Journal of Clinical Oncology, 7*, 145–151.

Schneider, J., Peltri, G., Bitterlich, N., Philipp, M., Velcovsky, H. G., Morr, H., et al. (2003). Fuzzy logic-based tumor marker profiles improved sensitivity of the detection of progression in small-cell lung cancer patients. *Clinical and Experimental Medicine, 2*, 185–191.

Schneider, J., Bitterlich, N., Kotschy-Lang, N., Raab, W., & Hans-Joachim, W. (2007). A fuzzy classifier using a marker panel for the detection of lung cancer in asbestosis patients. *Anticancer Research, 27*, 1869–1878.

Swensen, S. J., Jett, J. R., Hartman, T. E., Midthun, D. E., Mandrekar, S. J., Hillman, S. L., et al. (2005). CT screening for lung cancer: Five-year prospective experience. *Radiology, 235*, 259–265.

The International Early Lung Cancer Action Program Investigators. (2006). Survival of patients with stage I lung cancer detected on CT screening. *The New England Journal of Medicine, 355*, 1763–1771.

Travis, W. D., Brambilla, E., Mueller-Hermelink, H. K., & Harris, C. C. (2004). World Health Organization classification tumours of the lung, pleura, thymus and heart. *International Agency for Research on, Cancer*, 10–20.

Tureci, O., Mack, U., Luxemburger, U., Heinen, H., Krummenauer, F., Sester, M., et al. (2006). Humoral immune responses of lung cancer patients against tumor antigen NY-ESO-1. *Cancer Letters, 236*(1), 64–71.

Wu, Y., Wu, Y., Wang, J., Yan, Z., Qu, L., Xiang, B., et al. (2011). An optimal tumor marker group-coupled to artificial neural network for diagnosis of lung cancer. *Expert Systems with Applications, 38*, 11329–11334.

Zhang, Z., Zhou, H., Liu, S., & Harrington, P. B. (2006). An application of Takagi–Sugeno fuzzy system to the classification of cancer patients based on elemental contents in serum samples. *Chemometrics and Intelligent Laboratory Systems, 82*, 294–299.

Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine, 24*, 25–36.

Zhong, L., Coe, S. P., Stromberg, A. J., Khattar, N. H., Jett, J. R., & Hirschowitz, E. A. (2006). Profiling tumor-associated antibodies for early detection of non-small cell lung cancer. *Journal of Thoracic Oncology, 1*(6), 513–519.