VERSITA

Cent. Eur. J. Med. • 9(5) • 2014 • 632-641
DOI: 10.2478/s11536-013-0327-6

Central European Journal of **Medicine**

Artificial neural networks approach to early lung cancer detection

Research Article

Krzysztof Goryński*[1], Izabela Safian[2], Włodzimierz Grądzki[3], Michał Piotr Marszałł[1], Jerzy Krysiński[4], Sławomir Goryński[5], Anna Bitner[6], Jerzy Romaszko[7], Adam Buciński[2]

1 Department of Medicinal Chemistry, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, 85-089 Bydgoszcz, Poland

2 Department of Biopharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, 85-089 Bydgoszcz, Poland

3 Ward of Diagnostic-monitoring of Tuberculosis and Illness of Lungs, Voivodship Centre of the pulmonology, 85-326 Bydgoszcz, Poland

4 Department of Pharmaceutical Technology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, 85-089 Bydgoszcz, Poland

5 Department of Paliative Medicine, Regional Specialist Hospital in Grudziadz, 86-300 Grudziadz, Poland

6 Chair and Department of Hygiene and Epidemiology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, 85-094 Bydgoszcz, Poland

7 NZOZ Pantamed Sp z o.o. in Olsztyn, ul. Pana Tadeusza 6, 10-461, Olsztyn, Poland

Received 10 October 2012; Accepted 5 March 2014

Abstract: Lung cancer is rated with the highest incidence and mortality every year compared with other forms of cancer, therefore early detection and diagnosis is essential. Artificial Neural Networks (ANNs) are "artificial intelligence" software which have been used to assess a few prognostic situations. In this study, a database containing 193 patients from Diagnostic and Monitoring of Tuberculosis and Illness of Lungs Ward in Kuyavia and Pomerania Centre of the Pulmonology (Bydgoszcz, Poland) was analysed using ANNs. Each patient was described using 48 factors (i.e. age, sex, data of patient history, results from medical examinations etc.) and, as an output value, the expected presence of lung cancer was established. All 48 features were retrospectively collected and the database was divided into a training set (n=97), testing set (n=48) and a validating set (n=48). The best prediction score of the ANN model (MLP 48-9-2) was above 0.99 of the area under a receiver operator characteristic (ROC) curve. The ANNs were able to correctly classify 47 out of 48 test cases. These data suggest that Artificial Neural Networks can be used in prognosis of lung cancer and could help the physician in diagnosis of patients with the suspicion of lung cancer.

Keywords: *Artificial Neural Networks • Cancer diagnosis • Lung cancer • Risk factors*

1. Introduction

Lung cancer is a serious, worldwide health and epidemiological problem. It is defined as the uncontrolled growth of abnormal cells in the lung. In 2009, about 15% of all cancer cases and 29% of all cancer-related deaths were due to lung cancer [1,2]. It is rated with the highest death rate every year and is the second most diagnosed after

* *E-mail: gorynski@gmail.com*

prostate (28%) and breast (28%) cancers (for men and women, respectively) [1,2]. Lung cancer is most often found in elderly people because it develops over a long period of time. Falling ill with lung cancer peaks after the age of 70 years [3].

Many lung cancers are diagnosed at an advanced stage, leading to a poor prognosis. Early diagnosis of lung cancer is therefore important to facilitate treatment and potential cure whilst the disease is still in its early stages [4,5]. An Artificial Neural Network (ANN) is a tool that be used to assist the physician with the diagnosis of patients with the suspicion of the lung cancer.

An ANN is a computational model based on the human brain. ANNs contain some nodes which are connected through weights. Each node receives data from previous nodes, adds it together and outputs data through a nonlinear function, and then propagates data to proceeding nodes. The first neuronal layer of the ANN is the input layer composed of variable number of collected data from observation. The next additional neuronal layers compose of hidden layers created to generate a variable number of numerical combinations. The last neuronal layer named output layer generates the answer (numbers that represent the output). Most of ANN models are based on the idea of supervised training. There are two phases in the ANN action: training/learning phase and test phase. A validating phase is still often made. In training/learning phase, input data are presented to the ANN and weights are adjusted and fixed. In other words, the ANN does indeed learn the input patterns in the learning phase. In folllowing test phase, the data which are not used in the previous (training/learning) phase are presented to the ANN and the ANN's outputs are used to estimate its learning. If the estimate of ANN's action is satisfactory, it can be used in its own specific application [6-8].

Many scientists from different fields of technology and science use neural networks to solve problems in control, pattern classification, function approximation and medical diagnosis. In recent times ANNs are becoming quite popular in medicine, particularly for clinical diagnosis based on experimental and clinical data [9-13]. ANNs are a convenient tool in various tasks such as for blood cell classification, EEG, EMG, ECG analysis and bone fracture healing assessment, for diagnosis, healing, and prognosis in hypertension [14-16]. Recently many publications refer to the application of neural networks to predict and diagnosis breast, ovarian, gastrointestinal, bronchial, skin or prostate cancer [17-20]. More often the ANNs performance is compared with traditional screening methods using in oncology e.g. in prostate cancer (prostate specific antigen, PSA), serum marker, digital rectal examination, age and race, and Gleason sum. The published literature in the last decade suggests that ANN models have proven to be valuable tools in reducing the workload of oncologists by detecting cancer symptoms and providing decision support, potentially with the ability to automatically estimate the model on-line.

In this study, the ANN model was successful built and used to achieve a highly accurate set of patients with developing or already suffering from lung cancer from a wide, healthy population. In other words, ANNs can be a useful tool for the selection of the smallest group with highest risk of developing lung cancer.

2. Materials and methods

2.1. Patient characteristics

The study was based on retrospectively reviewed clinical records of patients treated at the Diagnostic and Monitoring of Tuberculosis and Illness of Lungs Ward in Kuyavia and Pomerania Centre of the pulmonology in Bydgoszcz (Poland) in 2006. The data collected included demographic factors (e.g. age, education, place of residence, etc.), clinical data from biochemical measurements (e.g. haemoglobin, creatinine, hematocrit, etc.), results obtained from spirometry tests (e.g. forced expiratory volume in 1 second, forced vital capacity, etc.) as well as information gathered from interview (e.g. lung cancer in the family, dyspnea, weight loss, etc.) and examination (e.g. auscultatory changes above pulmonary fields, form of the X-ray examination, etc.) done by primary care physician and specialist doctor. All variables considered in this study are presented in Table 1.

A total of 193 patients entered the study, aged between 26 and 85 years. Within the study group, the median age was 61 years, 70.5% were men, and more than 40% patients live in the provinces. The majority of the respondents were current cigarette smokers (91.7%). Diagnosis of lung cancer was reported for 103 patients. Baseline characteristics of the study population are shown in Table 2.

This work has been approved by the Research Ethics Committee at the Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, Poland.

2.2. ANN Analysis

Artificial neural networks were used for the analysis data and Statistica v. 8 software (StatSoft, Inc. Statsoft, Tulsa, OK., USA) was applied. Statistica Neural Networks software randomly divides data set into three

Table 1. Variables considered in the ANN analysis.

No	Variable name	Variable value
1	Age	years
2	Height	cm
3	Weight	kg
4	Sex	(1) – female, (2) - male
5	Zubrod score	(0) - asymptomatic (Fully active, able to carry on all predisease activities without restriction) (1) - symptomatic but completely ambulatory (Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature. For example, light housework, office work) (2) - symptomatic, <50% in bed during the day (Ambulatory and capable of all self care but unable to carry out any work activities. Up and about more than 50% of waking hours) (3) - symptomatic, >50% in bed, but not bedbound (Capable of only limited self-care, confined to bed or chair 50% or more of waking hours) (4) - bedbound (Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair)
6	Place of residence	(1) – capital of province, (2) – city of county, (3) – borough, (4) – village, (5) - homeless
7	Education	(0) – N/A, (1) – elementary, (2) – vocational, (3) – secondary, (4) - university
8	Occupational exposure	(0) – ND, (1) – absence of exposure, (2) – asbestos, (3) – silicas, (4) - aromatic organic compounds
9	Smoking cigarettes	(0) - non-smoker, (1) – smoker
10	Chronic obstructive pulmonary disease	(0) - lack of illness, 1 - illness is appearing
11	Past disease of pulmonary tuberculosis	(0) - lack of illness, (1) - illness is appearing
12	Other illness of lungs in the anamnesis	(0) - lack of illness, (1) - illness is appearing
13	Cancer of other organ	(0) - lack of illness, (1) - illness is appearing
14	Lung cancer in the family	(0) - lack of illness, (1) - illness is appearing
15	Pneumonia	(0) - lack of illness, (1) - illness is appearing
16	Cardiovascular diseases	(0) - lack of illness, (1) - illness is appearing
17	Other complication	(0) - lack of illness, (1) - illness is appearing
18	Cough	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
19	Dyspnea	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
20	Hemoptysis	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
21	Weight loss	numerical value [kg]
22	Lack of appetite	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
23	Pain complaints in the chest	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
24	Fever	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
25	Bone pains	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
26	Weakness	(0) - lack of symptom, (1) - current disease symptom in the anamnesis
27	Auscultatory changes above pulmonary fields	(0) - lack of changes, (1) – notice of changes
28	Presence of supraclavicular swollen glands or visible the bonfire smuggling in physical examining	(0) – lack, (1) – stated presence
29	Manifestations of syndrome of the vain vein upper or hoarseness above the month	(0) - lack of symptom, (1) - current of symptom
30	Neurological manifestations from the central nervous system	(0) - lack of symptom, (1) - current of symptom
31	Form of the X-ray examination	(0) - lack of visible changes, (1) – central form, (2) – peripheral form. (3) – scattered changes
32	Changes of the chest from the X-ray examination	(0) - lack of visible changes, (1) - widening the alcove, (2) - visible tuber without atelectasis of pulmonary pulp, (3) - presence atelectasis of the piece or the smaller area, (4) - atelectasis of the entire lung, (5) - liquid in the pleuritic hollow, (6) - high placing the diaphragm, (7) - the second bonfire or satellite nodules

Table 1 continued. Variables considered in the ANN analysis.

No	Variable name	Variable value
33	White Blood Cells (WBC)	numerical value [10^3/mL]
34	Red Blood Cells (RBC)	numerical value [10^6/mL]
35	Haemoglobin	numerical value [g/dL]
36	Hematocrit (HCT)	numerical value [%]
37	Platelets (PLT)	numerical value [10^3/mL]
38	Erythrocyte Sedimentation Rate	numerical value [10^3/mL]
39	Creatinine	numerical value [mg/dL]
40	Glutamic Oxoloacetic Transaminase (AspAT)	numerical value [IU/L]
41	Na	numerical value [mmol/L]
42	K	numerical value [mmol/L]
43	forced expiratory volume in 1 second (FEV1)	numerical value [L]
44	FEV1/FVC-FEV1-Percent (FEV1%) Forced Vital Capacity	numerical value [%]
45	Forced Vital Capacity (FVC Ex)	numerical value [L]
46	Forced Vital Capacity (FVC Ex%)	numerical value [%]
47	Vital Capacity (VC)	numerical value [L]
48	Vital Capacity (VC%)	numerical value [%]

N/A – not available; ND – no data

sections: a learning set, a validation set and a testing set with 97, 48 and 48 of cases respectively. Data of 193 patients was described by 48 parameters each of which was assigned a specific numerical value representing the feature of interest. Variables and their way of encoding them are presented in Table 1. Before commencing calculations, the whole data set was scaled up within the 0-1 range because variables with too great a value could mask out changes of importance. The output data were presented as two-state nominal variables–patients with lung cancer (encoding as 1) and without lung cancer (encoding as 0). Generally, an ANN, based on an multilayer perceptron (MLP) network consisting of one input layer, one hidden layer and one output layer, was used. This type of neural network is popular for classification and prediction application [9,13,20-22]. Initially, the different ANNs were designed and built to achieve the best network with optimum parameters. Next, the promising structures of the ANN were subjected to teaching 10 models for each comparison step. The most important factor in the MLP structure is the choice of the number of the hidden neurons, transfer function as well as training algorithm. The number of Processing Elements (PEs) in the hidden layer was changed from between 3 and 15, and three different transfer functions were tested: linear, tangent sigmoid and logistic sigmoid (Table 2). The weights were adapted during the training process using different back-propagation algorithms [23,24]: gradient descent, BFGS (Broyden-Fletcher-Goldfarb-Shanno) and Scaled Conjugate Gradient (Table 4), with the mean square error as the error measure. Basically, the weights are adjusted from cycle to cycle based on the information of gradient of the error function in the gradient method of learning. Algorithms used to optimize the learning coefficient are presented in equation 1.

$$w(k+1) = w(k) + \eta p(k) \qquad (Eq.1)$$

where η is the learning coefficient calculated in each cycle and $p(k)$ is the search direction vector of minimization the *kth* cycle. For BFGS algorithm, p(k) is defined as (Eq.2)

$$p(k) = -H^{-1}(k)g(k), \qquad (Eq.2)$$

where H(k) is an approximated Hessian matrix (second derivatives) and g(k) is a gradient vector of the error function in the kth learning cycle. Before the learning process, the network's weights were randomized with the value of the learning coefficient (η) = 0.1.

The optimal structure of the network that allowed for the correct classification of patients with or without lung cancer is presented in Figure 1. It consists of a Multilayer Perceptron with 48 input neurons (48 parameters describing 193 patients), 9 neurons in hidden layer (experimentally chosen) and 2 output layer (patients with lung cancer or without). Supervised method of learning with the BFGS algorithm (Broyden-Fletcher-Goldfarb-Shanno)

Table 2. Characteristics of the participants.

Characteristics	No of subjects (%)	No (%) of subjects with lung cancer
Total	193	103
Sex		
Female	57 (29.5)	28 (27.1)
Male	136 (70.5)	75 (72.9)
Age (median) range age	61 26 - 85	63 42 - 85
Place of residence		
Capital of the province	80 (41.5)	35 (33.9)
City of county	39 (20.2)	29 (28.2)
Borough	22 (11.4)	6 (5.8)
Village	51 (26.4)	32 (31.1)
Homeless	1 (0.5)	1 (1.0)
Education		
N/A	60 (31.1)	33 (32.1)
Elementary school	62 (32.1)	38 (36.9)
Vocational	37 (19.2)	20 (19.4)
High school	28 (14.5)	10 (9.7)
University	6 (3.1)	2 (1.9)
Smoking cigarette		
Yes	177 (91.7)	99 (96.1)
No	16 (8.3)	4 (3.9)

N/A: not applicable

process of teaching was applied. The activation function was used to scale the signal in the layer, and the hidden layer consisted of 9 artificial neurons with the linear activation function [Eq. (3)], and the output layer was formed by two neurons of hyperbolic tangent function (tanh) [Eq. (4)]. The network was found in the 17th cycle of teaching (Figure 2).

Eq. (3) ax.

Eq. (4) $\frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$, where *a* is the activation neuron (with the range -∞ to +∞).

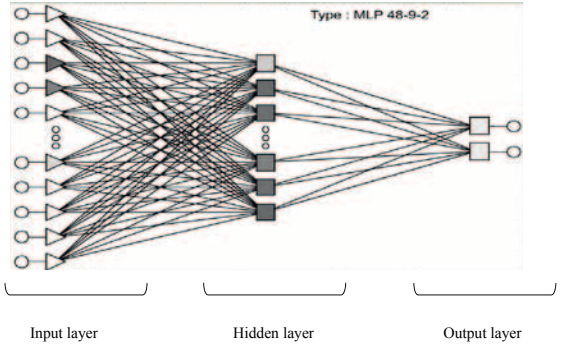

Figure 1. Architecture of the Artificial Neural Network applied.

3. Results

In this paper, by using MLP neural networks, a highly applicable method for accurate predicting patients with developing or already suffering lung cancer from healthy population has been presented. In order to find a final model many trials with different architecture has been performed. At the beginning, there were various choices

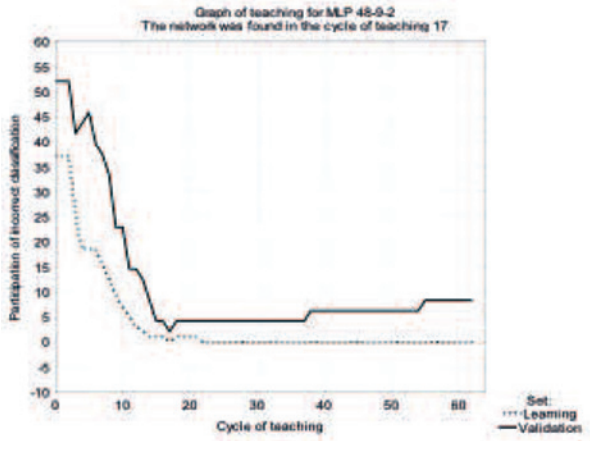

Figure 2. Graph of teaching for MLP 48-9-2

Table 3. A comparison of hidden and output activation function

No.	MLP network architecture	Performance (%)			Number of cycles in BFGS algorithm	Hidden activation function	Output activation function
		Training	Validation	Test			
1.	48-9-2	100	97.91	97.91	15	Linear	Tanh
2.	48-5-2	98.96	100	97.91	16	Linear	Tanh
3.	48-10-2	91.75	72.91	70.83	15	Linear	Log
4.	48-7-2	100	83.30	83.30	17	Log	Log
5.	48-15-2	100	95.83	93.75	17	Log	Tanh
6.	48-6-2	95.87	81.25	81.25	15	Tanh	Log
7.	48-15-2	100	97.91	97.91	30	Tanh	Linear
8.	48-4-2	100	97.91	93.75	20	Log	Linear
9.	48-14-2	100	97.91	91.66	21	Tanh	Tanh
10.	48-15-2	98.96	100	97.91	23	Linear	Tanh

MLP: Multilayer Perceptron Network; BFGS: Broyden-Fletcher-Goldfarb-Shanno training algorithm; Linear: activation level is passed on directly as the output of the neurons; Log: Logistic, S-shaped (sigmoid) curve, with output in the range (0, 1); Tanh: hyperbolic tangent function, S-shaped function, with output in the range (-1, +1).

for the transfer function in the hidden and output layers. The best results in terms of performance values in training, validation and test group, were obtained with networks using linear and tangent function in hidden and output layers, respectively (networks number 1, 2 and 10 from Table 3). Network number 7 with linear function in output layers also presented satisfactory performance values, however in the present study the output of the network has always two values, 0 (patients without lung cancer) or 1 (patients with lung cancer), so S-shaped (sigmoid) function (logistic or tangent) as output activation function should be selected. In order to define the number of neurons in the hidden layer, several trials with three different back propagation teaching algorithm were made. According to the comparison presented in Table 4, BFGS and conjugate gradient algorithm used for finding best possible weights values gave the best results in terms of performance and minimum number of training cycles, after that the program is stopped to prevent over-fitting problems. From the results achieved in the presented trials, two networks, MLR 48-9-2 with BFGS and MLR 48-15-2 with conjugated gradient applied algorithm, were selected because in preparing diagnoses are in 97.91% correct (47 in 48 test cases) and possess maximum values in learning and validation performance. Since MLR-48-9-2 takes into account the number of neurons in the hidden layer (lowest prevents over-fitting) and the number of training cycles (fewer use low computing memory) MLR 48-9-2 was the final model. Figure 2 illustrates the graph of teaching for ANN for the learning set (dotted line) and validation set (continuous line). The process of teaching was stopped after 17 cycles because further teaching did not lead to improvement, and after 17 cycles slightly over-fitting was observed. To the evaluate the performance of discrimination, in other words, the ability of the test to correctly classify those with and without the disease in question, the area under the curve (AUC) of the receiver operating curve (ROC) in training, validating and testing group (Table 5) was calculated. The ROC of our model (99.83%) implies perfect diagnosis.

The sensitivity analysis for the input variables was carried out and results are presented in Table 6. This procedure gives insight into the usefulness of individual variables. Variables with rank close to 1 are the most significant and with rank close to 48 are the least significant.

4. Discussion

According to sensitivity analysis, occupational exposure is the most significant parameter to the correct classification of patients with or without lung cancer. Long-term occupational exposure to carcinogenic industrial substances is the most common risk factor for lung cancer. Exposure to various carcinogens like arsenic, diesel exhaust, silica, asbestos, polycyclic hydrocarbons or beryllium has been linked to lung cancer [25,26]. In addition the incidence peak of lung cancer occurs 25-30 years after exposure [27]. Highly significant in our analysis are also red blood cells, Zubrod score, muscle weakness and smoking cigarettes. The high sensitivity coefficients obtained for red blood cells can help

Table 4. A comparison of performance (as a percentage of correct classifications) in found networks having different numbers of neurons in hidden layer using different training algorithm.

Number of neuron in hidden layer	Gradient descent Performance (%)				BFGS Performance (%)				Conjugate gradient Performance (%)			
	learning	validation	test	Number of cycles	learning	validation	test	Number of cycles	learning	validation	test	Number of cycles
3	86.59	87.50	91.66	180	100	100	95.83	20	100	97.91	97.91	20
4	86.59	93.75	93.75	182	100	97.91	97.91	18	100	97.91	97.91	28
5	86.59	91.66	91.66	178	100	97.91	95.83	19	100	97.91	97.91	22
6	86.59	91.66	91.66	179	100	97.91	97.91	19	100	97.91	97.91	23
7	86.59	93.75	91.66	179	100	97.91	97.91	14	100	97.91	97.91	27
8	86.59	93.75	91.66	179	100	97.91	95.83	21	100	97.91	97.91	23
9	86.59	87.50	91.66	180	100	100	97.91	17	100	97.91	97.91	30
10	86.59	91.66	91.66	177	100	97.91	95.83	31	100	97.91	97.91	23
11	85.56	93.75	91.66	178	100	95.83	95.83	18	100	97.91	97.91	23
12	85.56	93.75	91.66	178	100	100	95.83	30	100	97.91	97.91	27
13	86.59	93.75	91.66	178	100	97.91	97.91	20	100	97.91	97.91	32
14	95.56	93.75	91.66	177	100	97.91	97.91	15	100	97.91	97.91	31
15	86.59	89.58	91.66	178	98.96	100	97.91	13	100	100	97.91	25

Linear and hyperbolic tangent function were applied, in hidden and output layers, respectively.
BFGS: Broyden-Fletcher-Goldfarb-Shanno training algorithm

Table 5. The area under the ROC

No.	The area under the ROC
1.	Learning set 1.0000
2.	Testing set 0.9896
3.	Validating set 1.0000
4.	Learning, testing and validating set 0.9983

Table 6. Sensitivity analysis results for the variables used to estimate presence lung cancer in ANN analysis

Variable	Rank	Error
Occupational exposure	1	4.390
Red Blood Cells (RBC)	2	1.142
Zubrod score	3	1.107
Weakness	4	1.053
Smoking cigarettes	5	1.043
Erythrocyte Sedimentation Rate	6	1.034
Pain complaints in the chest	7	1.028
Pneumonia	8	1.024
Forced expiratory volume in 1 second (FEV1)	9	1.021
Other illness of lungs in the anamnesis	10	1.020
Hematocrit (HCT)	11	1.016
Haemoglobin	12	1.010
Weight loss	13	1.008
Na	14	1.007
Sex	15	1.007
Auscultatory changes above pulmonary fields	16	1.006
Forced Vital Capacity (FVC Ex)	17	1.006
White Blood Cells (WBC)	18	1.003
Cough	19	1.002
Lack of appetite	20	1.002
Presence of supraclavicular swollen glands or visible the bonfire smuggling in physical examining	21	1.001
Chronic obstructive pulmonary disease	22	1.001
Manifestations of syndrome of the vain vein upper or hoarseness above the month	23	1.001
Fever	24	1.000
Cardiovascular diseases	25	1.000
Other complication	26	1.000
Bone pains	27	0.999
K	28	0.999
Height	29	0.998
Age	30	0.998
Lung cancer in the family	31	0.998
Weight	32	0.996
Form of the X-ray examination	33	0.995
Glutamic Oxoloacetic Transaminase (AspAT)	34	0.994
Dyspnea	35	0.994
Platelets (PLT)	36	0.994
FEV1/FVC-FEV1-Percent (FEV1%) Forced Vital Capacity	37	0.994
Creatinine	38	0.993
Vital Capacity (VC)	39	0.991
Vital Capacity (VC%)	40	0.991
Cancer of other organ	41	0.990
Forced Vital Capacity (FVC Ex %)	42	0.987
Place of abode	43	0.984
Hemoptysis	44	0.982

determine progressive neoplasmatic diseases. The low erythrocyte level in serum is specific for most of the patients with advanced cancer disease, as an expression of secondary anemia. Zubrod score, also called ECOG scale (Eastern Cooperative Oncology Group efficiency scale) is a high sensitivity coefficient from the clinical point of view. ECOG scale allows to determine general condition and quality of life of patients with cancer (also used in other serious and chronic diseases); it is specified in degrees from 0 (normal performance) to 5 (death) describing the level of the patient's physical efficiency.

Cigarette smoking is the best-recognized and intensively described cause of lung cancer disease. Around 85-90% of lung cancer cases could be attributed to the use of tobacco indirectly or directly [28]. The risk of dying due to lung cancer increase 11-20 times compared with nonsmokers. The risk depends on the duration of smoking and the number of cigarettes smoked per day [29]. According to statistical data about 5% of all lung cancer cases constitutes second-hand smoking (about 25% of lung cancer in nonsmokers is assigned to second-hand smoking). The other relationship is that the lung cancer occurs most often in men than in women, which is closely following the past patterns of smoking prevalence [30]. In ranking sensitivity the last place, to our surprise, was 'Changes of the chest from the X-ray examination'. This is probably associated with the fact that potentially malignant cell clones are significantly below the limit of detection of current diagnostic tests [31], so could be not effective for early detection.

In summary, artificial neural networks are a class of discrimination statistical and nonlinear regression methods. They are quite popular in many areas of medicine. The improvement in prediction ability may be clinically important for clinical trials, therapy, and quality assurance. In decision-making concerning therapy, it may allow for the sufficient separation of patients with an excellent prognosis (who require little or no therapy) and those with a poor prognosis (who require therapy).

5. Conclusion

ANN can be used to estimate several prognostic situations, unlike other multiple statistical methods designed to calculate risk of diverse patient populations. In one of our laboratories we used the tool for estimation of the long-term prognosis of patients with lung cancer using multiple pathologic, clinical and laboratory features.

The proposed ANN model is efficient in predicting lung cancer risk. Hence, our model can improve the current diagnosis and prognosis methods. The main advantage of ANN is commercially available software which can be used in any desktop computer or standard laptop directly in the doctor's office. However, ANN will never replace expert opinion of the doctor but they can help in screening and can be used by experts to double-check their diagnosis. Diagnostic tests such as computed tomography (CT) or chest radiographs scans are not effective for early detection, therefore probably ANN can be useful in the initial diagnostic testing.

Acknowledgements

This work was supported by grant no. 23/2010 from Nicolaus Copernicus University in Torun, Poland.

Conflict of interest statement

Authors state no conflict of interests.

References

[1] Jemal A., Siegel R., Xu J., Ward E., Cancer Statistics 2010, Ca. Cancer. J. Clin., 2010, 60, 277-300
[2] Siegel R., Naishadham D., Jemal A., Cancer statistics, 2012, CA Cancer J Clin., 2012 62, 10-29
[3] Szczuka I., Roszkowski-Ślisz K., Lung cancer in Poland in 1970-2004, Pneumonol. Alergol. Pol., 2008, 76, 19-28
[4] Ahmed K., Emran A.A., Jesmin T., Early detection of lung cancer risk using data mining, Asian Pac. J.Cancer Prev., 2013, 14, 595-598
[5] Flores J. M., Herrera E., Leal G, Artificial Neural Network-Based Serum Biomarkers Analysis Improves Sensitivity in the Diagnosis of Lung Cancer, IFMBE Proceedings 2013, 33, 882-885
[6] Bishop C.M., Neural networks for pattern recognition, New York, NY: Oxford University Press 1995.
[7] Dayhoff J.E., DeLeo J.M., Artificial neural networks: opening the black box, Cancer, 2001, 91, 1615-1635
[8] Cross S.S., Harrison F.H., Kennedy R.L., Introduction to neural networks. Lancet, 1995, 346, 1075-1079
[9] Baxt W.G., Application of artificial neural networks to clinical medicine, Lancet, 1995, 346, 1135-1138

[10] Amato F., López A., Peña-Méndez E. M., Artificial neural networks in medical diagnosis, J. Appl. Biomed., 2013, 11, 47–58
[11] Alkim E., Gürbüz E., Kiliç E., A fast and adaptiveautomated disease diagnosis method with an innovative neural network model, Neur. Networks, 2012, 33, 88–96
[12] Atkov O., Gorokhova S., Sboev A., Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters, J. Cardiol., 2012, 59, 190–194
[13] Buciński A., Bączek T., Kaliszan R., Nasal A., Krysiński J., Załuski J., Artificial Neural Network Analysis of Patient and Treatment Variables as a Prognostic Tool in Breast Cancer after Mastectomy, Adv. Clin. Exp. Med., 2005, 14, 973–979
[14] Patel J.L., Goyal R.K., Application of artificial neural networks in medicinal science, Curr. Clin. Pharmacol., 2007, 2, 217-226
[15] Atkov O., Gorokhova S., Sboev A., Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters, J. Cardiol., 2012, 59, 190–194
[16] Uğuz H., A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases., J. Med. Syst., 2012, 36, 61–72
[17] Esteva H., Bellotti M., Marchevsky A.M., Neural networks for the estimation of prognosis in lung cancer, In: Naguib R.N., Sherbet G.V. eds. Artificial neural networks in cancer diagnosis, prognosis and patient management. Boca Raton: CRC Press, 2001: 29-37
[18] Bucinski A., Marszall M.P., Krysinski J., Lemieszek A., Zaluski A., Contribution of artificial intelligence to the knowledge of prognostic factors in Hodgkin's lymphoma, Eur. J. Cancer Prev., 2010, 19, 308-312
[19] Cinar M., Engin M., Egin E. Z., Atesci Y. Z., Early prostate cancer diagnosis by using artificial neural networks and support vector machines, Expert Systems with Applications, 2009, 36, 6357-6361
[20] Marszałł M.P., Krysiński J., Sroka W. D., Nyczak Z., Stefanowicz M., Waśniewski T., Romaszko J., Buciński A. ANN as a prognostic tool after treatment of non-seminoma testicular cancer, Central Eur. J. Med. 2012, 7, 672-679
[21] Bączek T., Buciński A., Ivanov A.R., Kaliszan R., Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics, Anal. Chem., 2004, 76, 1726-1732
[22] Koba M., Application of artificial neural networks for the prediction of antitumor activity of a series of acridinone derivatives, Med. Chem., 2012, 8, 309-319
[23] Luenberger D.G., Ye Y., Linear and nonlinear programming, International Series in Operations Research & Management Science 116 (Third ed.), New York, Springer, 2008, pp. xiv+546
[24] Knyazev, A.V., Lashuk I.,Steepest Descent and Conjugate Gradient Methods with Variable Preconditioning, SIAM J. Matrix Anal. A, 2008, 29, 1267-1280
[25] Chen T.M., Kuschner W.G., Non-tobacco related lung carcinogens. Lung cancer principle and practice, In: Harvey P. et al, editors. 3rd ed, Lippincot Williams and Wilkins: 2005. p. 61-73
[26] Bij S., Hendrik Koffijberg H., Lenters V., Lung cancer risk at low cumulative asbestos exposure: meta-regression of the exposure–response relationship, Cancer Cause. Control, 2013, 24, 1-12
[27] Alberg A.J., Samet J.M., Epidemiology of lung cancer, Chest, 2003, 123, 21S-49S
[28] Shopland D., Tobacco use and its contribution to early cancer mortality with a special emphasis on cigarette smoking, Environ. Health. Prospect., 1995, 103, 131-142
[29] Doll R., Peto R., Boreham J., Sutherland I., Mortality from cancer in relation to smoking: 50 years observations on British doctors, Br. J. Cancer, 2005, 92, 426-429
[30] De Matteis S., Consonni D., Pesatori A.C., Are women who smoke at higher risk for lung cancer than men who smoke? Am. J. Epidemiol., 2013, 177, 601-612
[31] Sutedja G., New techniques for early detection of lung cancer, Eur. Respir. J., 2003, 21, 57–66