Ultrasound Obstet Gynecol 1999;13:17–25

98-45

Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses

D. Timmerman, H. Verrelst, T. H. Bourne†, B. De Moor*, W. P. Collins‡, I. Vergote and J. Vandewalle**

Department of Obstetrics and Gynecology, University Hospitals Leuven, Belgium; *Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium; †Academic Department of Obstetrics and Gynaecology, St. George's Hospital Medical School, London, UK; and ‡Department of Obstetrics and Gynaecology, Guy's, King's and St. Thomas' School of Medicine, King's College Hospital, London, UK

Key words: ULTRASONOGRAPHY, COLOR DOPPLER IMAGING, CA 125, NEURAL NETWORKS, OVARIAN CARCINOMA

ABSTRACT

Objective *The aim of this study was to generate and evaluate artificial neural network (ANN) models from simple clinical and ultrasound-derived criteria to predict whether or not an adnexal mass will have histological evidence of malignancy.*

Design *The data were collected prospectively from 173 consecutive patients who were scheduled to undergo surgical investigations at the University Hospitals, Leuven, between August 1994 and August 1996. The outcome measure was the histological classification of excised tissues as malignant (including borderline) or benign.*

Methods *Age, menopausal status and serum CA 125 levels and sonographic features of the adnexal mass were encoded as variables. The ANNs were trained on a randomly selected set of 116 patient records and tested on the remainder (n = 57). The performance of each model was evaluated using receiver operating characteristic (ROC) curves and compared with corresponding data from an established risk of malignancy index (RMI) and a logistic regression model.*

Results *There were 124 benign masses, five of borderline malignancy and 44 invasive cancers (of which 29% were metastatic); 37% of patients with a malignant or borderline tumor had stage I disease. The best ANN gave an area under the ROC curve of 0.979 for the whole dataset, a sensitivity of 95.9% and specificity of 93.5%. The corresponding values for the RMI were 0.882, 67.3% and 91.1%, and for the logistic regression model 0.956, 95.9% and 85.5%, respectively.*

Conclusion *An ANN can be trained to provide clinically accurate information, on whether or not an adnexal mass is malignant, from the patient's menopausal status, serum CA 125 levels, and some simple ultrasonographic criteria.*

INTRODUCTION

Adnexal masses are a very common problem in gynecology. Since certainty about the nature of an adnexal mass can only be gained after histological examination, the traditional strategy for establishing the final diagnosis has been to perform an exploratory laparotomy. However, many surgeons are unexpectedly faced with ovarian cancer at the time of surgery. Therefore, women with advanced ovarian cancer often undergo insufficient primary surgery at local hospitals[1] or in the hands of surgeons without appropriate skills and experience of debulking surgery[2]. Insufficient primary surgery leads to under-staging of patients with ovarian cancer and to poorer survival rates, because the amount of residual malignant tissue is one of the most important prognostic factors[2–5]. The timely referral of patients with a suspicious adnexal mass to an oncological surgeon for the appropriate pre-, per-, and postoperative management is therefore of utmost importance in order to increase survival rates. The debate regarding the preoperative characterization of persistent adnexal masses continues[6–10], and it is an issue of considerable importance in view of the increasing interest in the conservative management of ovarian tumors, as well as in selecting patients for less invasive surgical techniques[11,12].

Correspondence: Dr D. Timmerman, Department of Obstetrics and Gynecology, University Hospitals Leuven, Herestraat 49, B-3000 Leuven, Belgium

Received 28–8–98
Revised 1–12–98
Accepted 8–12–98

Previous studies have investigated the use of morphological scoring systems and multiple logistic regression[6,10,13–15]. As these methods do not (systematically) investigate the possible existence of non-linear interaction/correlation between variables, we applied more complex modelling techniques in this study. An increasingly popular technique – which has shown its robust applicability to non-linear regression and classification tasks – is artificial neural networking. This approach was developed in the artificial intelligence research community to very crudely mimic the workings of the human brain on a computer. The networks resemble the brain in two respects[16]: first, knowledge is acquired by the network through a learning process (i.e. training) and, second, interneuron connection strengths (known as synaptic weights) are used to store the knowledge. The main differences between an artificial neural network (ANN) and the brain are the form and speed of interactions between the neurons and the available number of neurons. Neurons in the brain are much slower (milliseconds) than silicon logic gates (nanoseconds) for the transmission of information, because they use chemical synapses. However, the estimated number of neurons in the human cortex (10 billion) and the number of synapses between them (60 trillion)[17] by far exceed the number of possible 'neurons' or 'processing units' in an ANN, which consists of a large interconnection of these simple computing cells.

Artificial neural networks can be described as networks of neurons that exchange information in the form of numerical values with each other via synaptic interconnections (see Figure 1). The neurons take a weighted sum of their stimuli (or inputs), subtract a certain inhibition bias (or threshold), and then put this numerical value through a saturation characteristic (or activation function) to produce an excitation level (or output). By efficiently interconnecting such neurons, powerful and flexible information processing structures arise. This determination of the optimal interconnections is referred to as the learning algorithm for the neural network. This algorithm modifies the synaptic weights of the network in an orderly fashion so as to attain the desired objective. A distinction is made between supervised and unsupervised networks. Unsupervised or

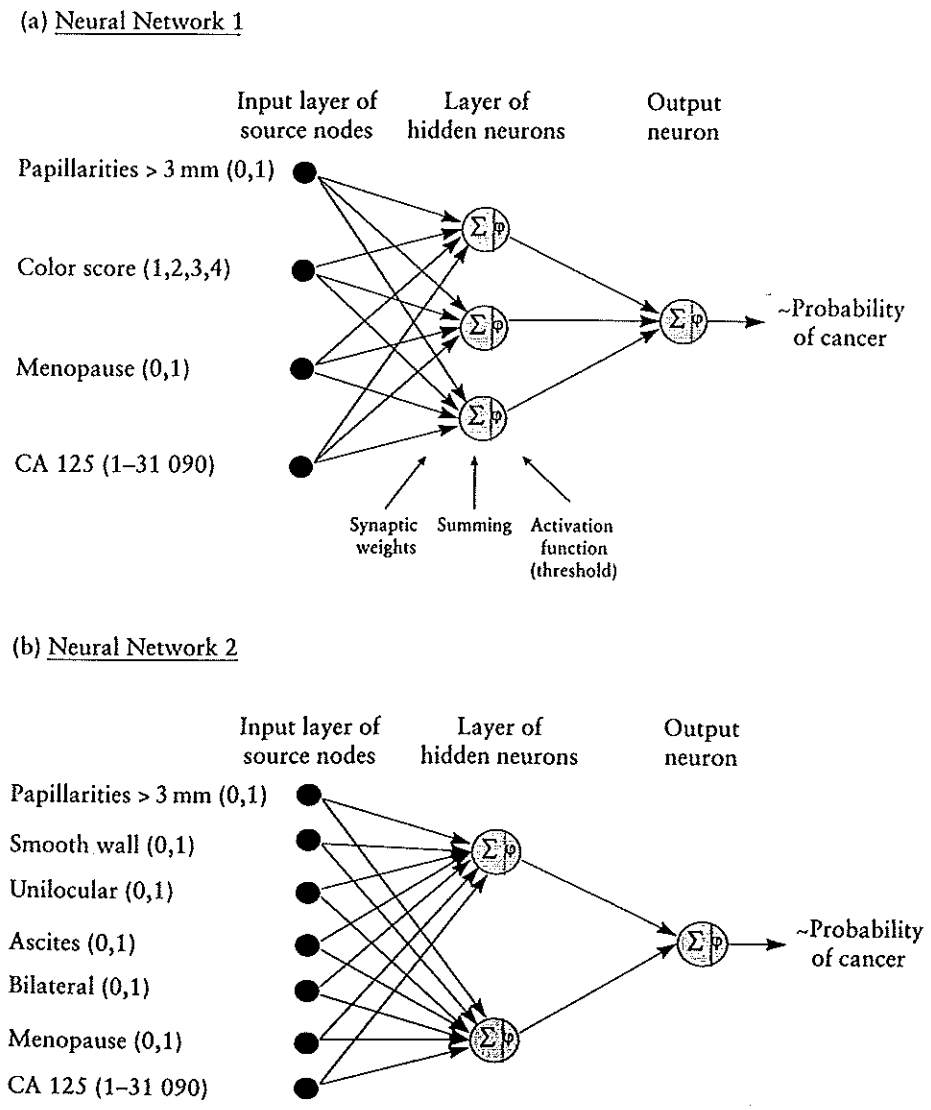


Figure 1 Architectural graph of Neural Network 1 (a) and 2 (b) which both represent multilayer feed-forward networks, containing one hidden layer

self-organizing networks are capable of partitioning an input feature set (e.g. morphological variables) into distinct subsets or classes with no *a priori* knowledge of subset classes. This means that the network decides which feature sets are similar[18]. On the other hand, supervised networks (the type of networks that are used in this study) require an external teacher to decide for the network which feature sets belong to the same class. The adjustment of the network weights (for example, through the back-propagation algorithm) and thus of the network's behavior is based on an external supply of examples of correct performance. If enough examples are available and if appropriate learning methods are used, the output of the overall neural network can be trained to approximate P (mass = $M \mid x_1, \ldots, x_n$), i.e. the post-test probability of a certain adnexal mass being malignant, given a number of clinical and ultrasonographic findings, encoded in the numerical values x_1–x_n.

The specific type of ANNs that will be considered here are called feed-forward multilayer perceptrons. As illustrated in Figure 1, these networks have a layered architecture: the information enters the network in the input layer, is fed forward into the hidden neuron layers where it is processed to result in the output layer. From the perspective of a statistician or a mathematician, feed-forward artificial neural networks can be classified as a non-linear extension to logistic regression modelling.

In logistic regression analysis, a weighted sum of measurement values $x_1, x_2, \ldots, x_n$ minus an intercept produces the logit $z = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n - \beta_0$. After applying the logistic function to this logit, we obtain $P = 1/(1 + exp(-z))$ as an estimate of the post-test probability of malignancy. One can see that a single neuron in the neural network operates essentially in the same way as a logistic regression model. The difference is that, in the feed-forward neural network, there are several such models in parallel (the connections between input and hidden layer), which are again weighted and summed together (the connections between the hidden layer and the output) to produce the probability of malignancy. It is this increase in complexity which allows the modelling of non-linear correlations between variables.

The aim of this study is to demonstrate how, through the use of artificial neural networks, statistical models can be obtained that combine the predictive value of single morphological or color Doppler parameters of the adnexal mass, the menopausal status of the patient and the level of serum tumor marker CA 125 in an optimal way. The results are compared with those obtained from an established risk of malignancy index (RMI)[13] and with a logistic regression model.

METHODS

We used transvaginal ultrasonography with color Doppler imaging to examine all patients referred for a preoperative assessment of an adnexal mass. The histological classification of the adnexal mass was used as the end-point for the study. The study group consisted of all consecutive patients who were referred to a single institution (University Hospitals Leuven) from August 1994 to August 1996[10]. Patients without a preoperatively measured serum CA 125 level were excluded from the study. Hysterectomized patients, older than 50 years of age, were classified as postmenopausal.

Ultrasonography

All patients were scanned by the same ultrasonographer (D.T.) with an Acuson 128 XP/10 Acoustic Response Technology (ART) ultrasound system (Acuson Inc., Mountain View, CA, USA) equipped for color Doppler imaging. A 5-MHz intravaginal probe was used, incorporating a field of view of 90°. The wall filter was set at 50 Hz and the pulsed Doppler sample volume size at 2 mm. An extensive gray-scale examination with attention to morphology was performed. Ascites was classified as present if there were at least 5 mm of free fluid in the pouch of Douglas. A septum was defined as a thin strand of tissue running across the cyst cavity from one internal surface to the contralateral side; it was classified as being less than, or equal to or greater than 3.0 mm in thickness. Solid papillary projections were structures that had penetrated into the cyst cavity and were more than 3.0 mm in length. Locularity was expressed as unilocular (a unilocular cyst without septa and without solid parts), unilocular–solid, multilocular, multilocular–solid or solid (a tumor where the solid components comprised 80% or more of the tumor)[14]. In cases of solid tumors, the description of the internal wall as smooth or irregular was not usually applicable. If separate cysts were visualized within one adnexal mass, it was described as multicystic. The echogenicity of the cystic contents was described as echolucent, low-level echogenic, 'ground glass' appearance (homogeneously dispersed echogenic cystic contents, as often seen in endometriotic cysts), hemorrhagic (with internal thread-like structures, representing fibrin strands) or mixed echogenic (as often seen in teratomas). The presence of acoustic shadows, defined as loss of acoustic echo behind a sound-absorbing structure, was also recorded.

Color Doppler imaging

Subsequently, the entire tumor was surveyed by color Doppler imaging. The power, gain and pulse repetition frequency were adjusted for maximum sensitivity of low blood flow states. The lowest velocity signals were filtered out by gradually increasing the pulse repetition frequency, and flow analysis was concentrated on the highest velocity signals. A pulsed Doppler range gate was placed over these areas to obtain flow velocity waveforms. A subjective semiquantitative assessment of the amount of blood flow within the septa, cyst walls, solid tumor areas or ovarian stroma was made: a score of 1 was recorded when no arterial blood flow could be found; 2 when only minimal flow (area and intensity of color) could be detected; 3 when moderately strong flow was present; and 4 when the adnexal mass contained very strong blood flow. Multiple

photographic prints were made of relevant structures and Doppler signals.

Serum CA 125 levels

At the request of the referring clinician, serum CA 125 levels had been measured preoperatively in the majority of patients. The CA 125 II immunoradiometric assay (Centocor, Malvern, PA, USA) was used and the ultrasonographer was unaware of the results at the time of the scan.

Staging

Malignancies were staged by a gynecological oncologist according to criteria recommended by the International Federation of Gynecology and Obstetrics (FIGO)[19].

Artificial neural network

All simulations and calculations in the development and analysis of the artificial neural networks and all calculations of receiver operating characteristic (ROC) curves, the area under these curves and of standard errors on these areas, were programmed in Matlab Version 5.0.0 and performed on a UNIX Sparc Ultra workstation. The dedicated programs were developed at the Signals, Identification, System Theory and Automatization (SISTA) laboratory of the Department of Electrical Engineering at the Katholieke Universiteit Leuven. Two one-hidden-layer feed-forward neural networks were considered. To avoid too many free variables in the network, which would enhance the possibility of overtraining and thus of bad generalization, only a limited number of input parameters and few hidden neurons were selected. We first created a simple ANN model (Neural Network 1) based on the same input variables as the logistic regression model proposed previously[10] (see Figure 1). This ANN consisted of four input neurons, one layer with three hidden neurons and one output. The inputs to the first network were the menopausal status (0 or 1), serum CA 125 level, the color score for tumor vascularity (1, 2, 3 or 4) and the presence or absence of papillary structures (0 or 1).

Subsequently, we tried to improve the performance of this model by changing the architecture and the input variables. Multivariate logistic regression analysis was used to make an initial subselection of relevant features. The training set of patient records was used to produce about 100 different ANN models, in which the architecture and inclusion of input features were varied, based on trial and error. Afterwards, these models were validated using the test set of patient records. A manual selection of the best performing model, using the training and the test set, was based on the values for sensitivity and specificity and the area under the resulting ROC curve. The seven selected inputs to this second ANN model (Neural Network 2) were menopausal status, the serum CA 125 levels and a set of morphological parameters: presence of ascites, unilocularity, smoothness of the internal wall, presence of papillary structures and whether the mass was bilateral. The numerical values to the inputs of the network were 0 or 1, relating to the observed presence or absence of each parameter. Only two hidden neurons were taken and one output. In both networks, the hidden neurons and the output neurons had the logistic function $1/(1 + \exp(-z))$ as saturation characteristics, where exp () is the natural exponent and z is the weighted sum of values going into the neuron. Training consisted of the parametric minimization of the mean squared error between desired and actual response of the neural network. To bias the performance towards good performance for the detection of malignancy (high sensitivity), the contribution of the malignant examples was given a double weight in this error. A quasi-Newton optimization method with the Broyden–Fletcher–Goldfarb–Shanno update of the Hessian ('fminu' routine in Matlab) was used to minimize the mean squared error. At the start of training, the weights and biases of the network were initialized with zero mean normally distributed random variables. Since a Newton method is a scale invariant technique, no standardization of the input values was needed. To avoid overtraining (i.e. the network being too much specialized in the training set), the optimization was stopped when the corresponding error on the test set started to increase again instead of decreasing, together with the error on the training set.

For reasons of comparison, we also developed a logistic regression model (LR2) based on the same input features as Neural Network 2. This training was done with the same subdivision in training and test set and with the same parametric optimization algorithm as used for NN2.

Statistical analysis

The predictions for both networks (with the final weight and bias values) and the logistic regression model were calculated for the training and test set. These values were used to construct ROC curves by varying the decision value between 0 and 1, with steps of 0.001. The area under each curve was approximated by trapezoidal integration. The standard errors (SE) associated with these areas were estimated by the method of Wilcoxon[20]. These SE were used for the comparison of the areas under two ROC curves by the method given by Hanley and McNeil[21]. The Pearson product-moment correlation method was chosen to calculate the correlation r between the areas. With this value, the SE on the difference in areas was calculated and a two-tailed test was performed on this difference by assuming Gaussian distributions.

The performance of the network models was also evaluated in terms of sensitivity, specificity, positive predictive value and negative predictive value over a range of decision levels.

RESULTS

A total of 191 consecutive patients were enrolled. However, 18 patients were excluded from the analysis because the level of serum CA 125 had not been measured preoperatively. Some characteristics of the study group of 173 patients are summarized in Table 1. Eighty-three

Table 1 Characteristics of patients by type of adnexal mass

Adnexal mass type	*Number of patients*	*Age (years)*		*% postmenopausal*	*% with serum CA 125 ≥ 35 U/ml*
		Mean	*Range*		
Primary invasive*	31	60	36–80	81	84
Metastatic invasive	13	64	40–86	85	67
Borderline†	5	38	27–49	0	60
Benign	124	51	22–93	44	19
All	173	53	22–93	52	36

*Of which 14 (45%) and †5 (100%) were stage I

women were premenopausal (mean age 40 years, range 22–57 years) and 90 were postmenopausal (mean age 65 years, range 47–93 years). Thirty-two women had bilateral adnexal masses. They only contributed data from a single tumor to the study. The histological classification of the worst case mass from each patient was used as the outcome measure for the study. Borderline malignant tumors were categorized with the malignant masses. None of the women had a benign tumor on one side and a malignant tumor on the other.

Histology and staging

At surgery, benign masses were found in 124 patients, borderline malignant masses in five patients and invasive cancers in 44 patients (see Table 2). In this referred group of patients, the prevalence of malignant and borderline malignant tumors was 28%. All borderline malignant masses were found in premenopausal patients (mean age of 38 years), whereas patients presenting with invasive tumors were postmenopausal in 81% of the cases (mean age of 60 years). The most common benign tumors were cystadenomas and endometriotic cysts. A high proportion (29%) of patients presenting with invasive adnexal malignancies had a metastatic tumor, commonly from gastrointestinal origin. Overall, 37% of patients presenting with a malignant or borderline malignant tumor had stage I disease.

Table 2 Histological classification of adnexal masses

Histological classification	*Patients*	
	n	%
Primary invasive		
Serous cystadenocarcinoma	18	58
Endometrioid cystadenocarcinoma	6	19
Transitional cell carcinoma	2	6
Mucinous cystadenocarcinoma	1	3
Clear cell cystadenocarcinoma	1	3
Granulosa cell tumor	1	3
Sarcoma	1	3
Poorly differentiated adenocarcinoma	1	3
Metastatic invasive		
Colon	4	31
Stomach	3	15
Endometrium	2	15
Sarcoma	2	15
Breast	1	8
Unknown	1	8
Borderline		
Serous cystadenocarcinoma	5	100
Benign		
Cystadenoma	44	36
Endometrioma	15	12
Teratoma	8	6
Follicular cyst	10	8
Cystadenofibroma	9	7
Corpus luteum cyst	6	5
Others	32	26

Artificial neural networks

From both classes (124 benign and 49 malignant adnexal masses), two-thirds of the examples were randomly selected to form the training set ($n = 83$ benign and 33 malignant), and the remaining one-third formed the test set ($n = 41$ benign and 16 malignant). For the network output to produce the probability of malignancy, the desired responses for benign and malignant examples were labelled 0 and 1, respectively. The architectures of the two networks are given in Figure 1. The performances of the selected models in comparison with the RMI, serum CA 125 level and age are presented in the ROC curves plotted in Figure 2.

In the subsequent paragraphs, the variable 'meno' stands for the patient being pre- or postmenopausal (values 0 or 1), 'CA 125' is the CA 125 level in U/ml. The variable 'color-score' is the subjective assessment of the amount of blood flow (1, 2, 3 or 4). 'Papill' stands for the presence or absence (1/0) of papillary projections larger than 3 mm. The variables 'ascites', 'smooth', 'unilocular' and 'bilateral', respectively denote whether or not (1/0) ascites was present, whether the adnexal mass had smooth internal walls, was unilocular and was present on both pelvic sides. The notation exp(z) stands for taking the natural exponent of the value z.

First network NN1

Training of the first network was stopped after 350 iterations when the mean squared errors on the training and test sets were 3.0% and 4.1%, respectively. The mathematical formula to calculate the prediction values of this neural network is:

$$Y = 1/(1 + \exp(z_y)), \text{ with } z_y = 1.9384h_1 + 3.2379h_2 + 3.3631h_3 - 5.4257, \text{ in which}$$

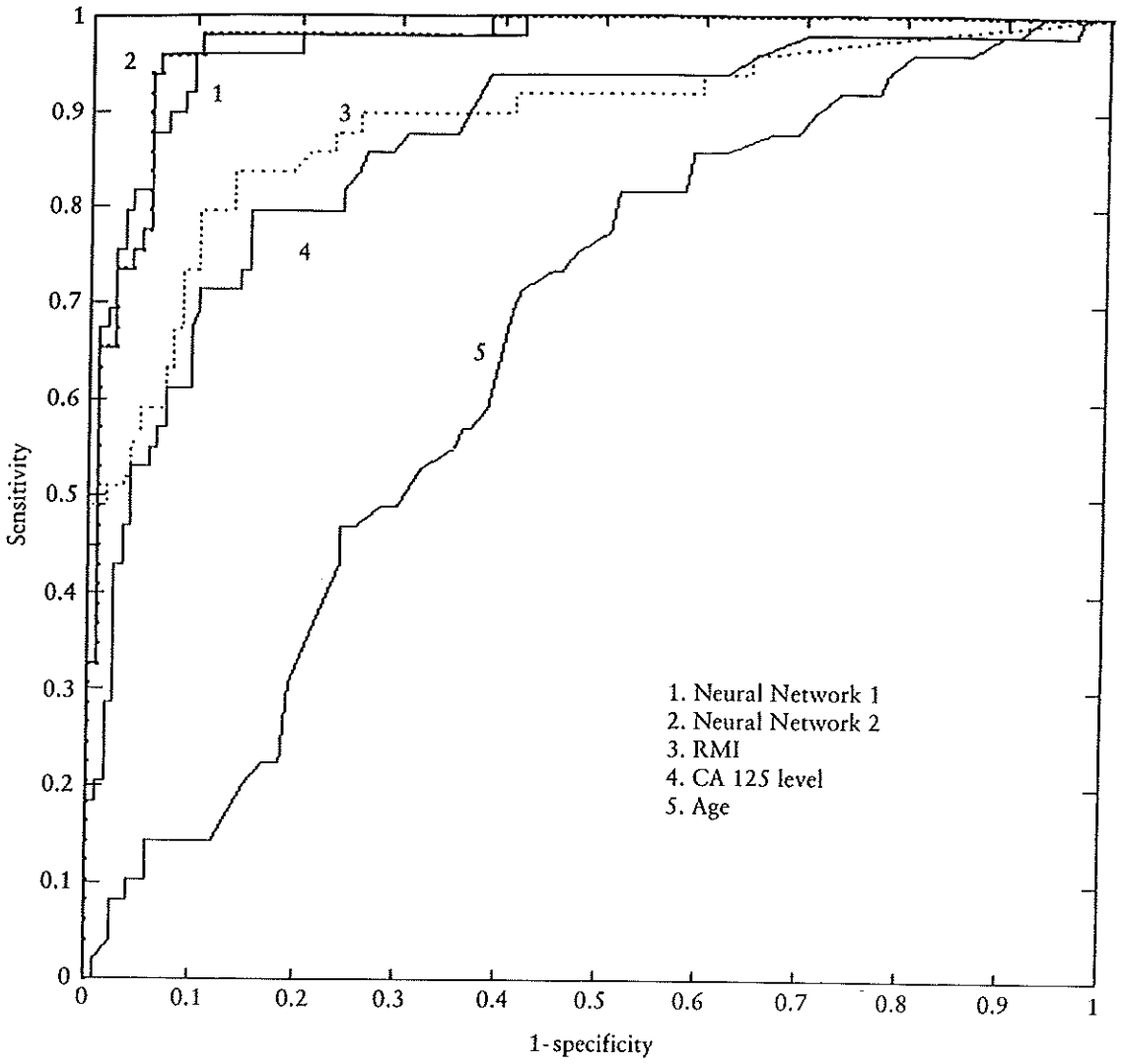


Figure 2 Malignant versus benign adnexal masses. ROC curves for tests based on patient's age, serum CA 125 level, the risk of malignancy index (RMI)[13] and two neural network models (training + test data)

$h_1 = 1/(1 + \exp(z_1))$, with $z_1 = -0.2425\text{meno} + 0.0782\text{CA 125} - 0.0381\text{color-score} + 0.8974\text{papill} - 1.2029$;

$h_2 = 1/(1 + \exp(z_2))$, with $z_2 = 0.2636\text{meno} + 0.0106\text{CA 125} + 0.4893\text{color-score} + 1.7447\text{papill} - 2.2744$;

$h_3 = 1/(1 + \exp(z_3))$, with $z_3 = 0.3075\text{meno} + 0.0001\text{CA 125} + 0.5429\text{color-score} + 1.3764\text{papill} - 1.8158$.

At a probability of malignancy greater than 45%, the first network gave a sensitivity of 97.0% and a specificity of 89.2% on the training set. When cross-validated on the test set, the use of this decision level gave a sensitivity and a specificity of 87.5% and 92.7%, respectively. The areas under the ROC curves were 98.0 ± 1.7%, 95.1 ± 3.9% and 97.1 ± 1.7% for training, test and complete set, respectively (Table 3). The first neural network performed significantly better than the risk of malignancy index ($p = 0.009$) and CA 125 levels ($p = 0.004$).

Second network NN2

Training of the second network was stopped after 300 iterations. The mean squared error was 2.1% and 3.2% on the training and test set, respectively. The mathematical formula of this network is:

$Y = 1/(1 + \exp(z_y))$, with $z_y = 2.9753h_1 + 4.1980h_2 - 3.8616$, in which

$h_1 = 1/(1 + \exp(z_1))$, with $z_1 = -1.0792\text{meno} + 1.9383\text{CA 125} + 0.7124\text{ascites} - 1.2664\text{unilocular} - 1.3741\text{smooth} + 0.8298\text{papill} + 1.5316\text{bilateral} - 0.5485$;

$h_2 = 1/(1 + \exp(z_2))$, with $z_2 = 1.0766\text{meno} + 0.1376\text{CA 125} + 1.0112\text{ascites} - 0.8320\text{unilocular} - 1.6941\text{smooth} + 2.9541\text{papill} + 1.4654\text{bilateral} - 1.8129$

At a probability of malignancy greater than 60%, the second network gave a sensitivity of 97.0% and a specificity of 92.8% on the training set. Cross-validated on the test set, this decision level gave a sensitivity and a specificity of 93.8% and 95.1%, respectively. This neural network also

Table 3 Comparison of the diagnostic performance of two neural network models used to discriminate between malignant and benign adnexal masses (n = 173)

Model	*n*	*Area under the ROC curve*	*SE*	*Best decision level*	*Sensitivity (%)*	*Specificity (%)*	*PPV (%)*	*NPV (%)*
Serum CA 125 (U/ml)		0.872	0.031	25	85.7	73.4	56.0	92.9
		0.872	0.031	35	79.6	81.5	62.9	91.0
Risk of malignancy index[13]		0.882	0.035	50	87.8	74.2	57.3	93.9
		0.882	0.035	100	83.7	82.2	65.1	92.7
		0.882	0.035	200	67.3	91.1	75.0	87.6
Neural Network 1 (training set)	116	0.980	0.017	45%	97.0	89.2	78.1	98.7
Neural Network 1 (testing set)	57	0.951	0.039	45%	87.5	92.7	82.4	95.0
Neural Network 1 (all patients)	173	0.971	0.017	45%	93.9	90.3	79.3	97.4
Neural Network 2 (training set)	116	0.976	0.019	60%	97.0	92.8	84.2	98.7
Neural Network 2 (testing set)	57	0.983	0.023	60%	93.8	95.1	88.2	97.5
Neural Network 2 (all patients)	173	0.979	0.015	60%	95.9	93.5	85.4	98.3
Logistic Regression 2 (training set)	116	0.951	0.027	60%	97.0	85.5	72.7	98.7
Logistic Regression 2 (testing set)	57	0.965	0.033	60%	93.8	83.0	68.2	97.1
Logistic Regression 2 (all patients)	173	0.956	0.021	60%	95.9	85.5	72.3	98.1

PPV, positive predictive value; NPV, negative predictive value

performed significantly better than the risk of malignancy index (p = 0.015) and CA 125 levels (p = 0.006). The areas under the ROC curves were 97.6 ± 1.9%, 98.3 ± 2.3% and 97.9 ± 1.5% for training, test and complete set, respectively (Table 3).

Logistic regression model LR2

The probability of malignancy predicted by the logistic regression model (based on the same inputs as NN2) is given by $y = 1/(1 + \exp(-z))$, with z = 0.5948meno + 0.0205CA 125 + 0.5446ascites − 0.7620unilocular − 1.1606smooth + 1.5049papill + 0.7633bilateral − 1.0889. At a probability of malignancy greater than 60%, this model gave a sensitivity of 97.0% and a specificity of 85.5% on the training set. Cross-validated on the test set, this decision level gave a sensitivity and a specificity of 93.8% and 83.0%, respectively. The areas under the ROC curves were 95.1 ± 2.7%, 96.5 ± 3.3% and 95.6 ± 2.1% for training, test and complete set, respectively (Table 3). The significance level of the difference between NN2 and the logistic regression model is p = 0.07. Figure 3 shows histograms of the predicted values for malignant and benign subpopulations to illustrate the performance of the test based on serum CA 125 levels, the logistic regression model and neural network NN2.

DISCUSSION

The design of a neural network and the selection of the optimal number of source nodes in the input layer and the optimal number of hidden neurons are difficult and result from trial and error, since no strict rules can be applied. A general rule is that a network, which is too complex for the available number of examples (i.e. size of the training set), will easily become overtrained (over-fitting) and therefore will have a poor generalization capacity. It will not know how to handle new cases (test data), which are different from the training data. However, if the network is not complex enough, it will not have a good learning capacity and the test results will be equally suboptimal. Therefore, it is vital to provide enough learning examples for any ANN in order to exploit its full potential.

The performance results of the training and the test sets of both neural networks are very encouraging. However, since the test results have been used to select the optimal ANN, they cannot be used for a prospective evaluation of the network's performance. Thus, these results have been obtained retrospectively, and they need to be cross-validated in a prospective trial. It is well known and theoretically studied that, due to the non-linear character of the neurons, the mean squared error is not a well-consorted criterion in the sense that (Newton) optimization methods from different initializations almost always converge to different end-results. This is known in neural network and optimization literature as the problem of local minima. However, even being suboptimal, the networks do have a small mean squared error and, when the areas under the ROC curves are compared, we see that they have a performance which is significantly better than that of single parameters such as the serum CA 125 level (area of 0.872) and the commonly used risk of malignancy index[13] (area of 0.882).

The difference in area under the ROC curve for the second neural network with a logistic regression model on the same features was close to significant (p = 0.07). To approximate the area under the ROC curve, we used the trapezoidal integration rule. The Wilcoxon statistic provided an estimate of the error on this area. Both these methods are considered to be conservative (area underestimated, error overestimated). Using other methods, more accurate assessments can be obtained on the significance of the difference in performance. Another way to enlarge our confidence in the significance would be to collect an (even) larger set of examples, as the error on the area is inversely proportional to the square of the number of examples in

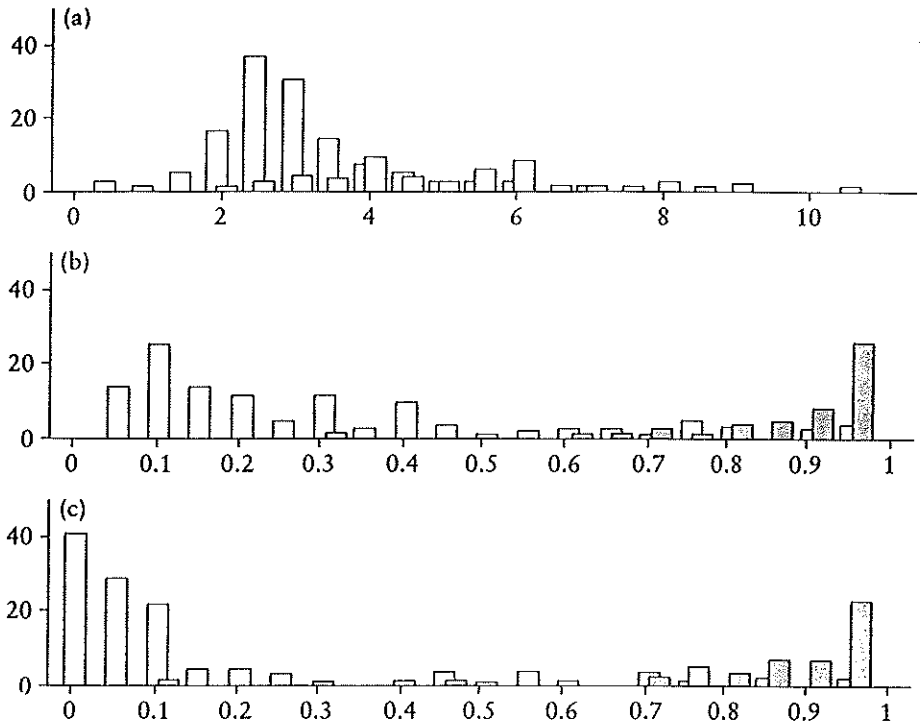


Figure 3 Histograms over benign (white) and malignant (gray) subpopulations of the predicted values of a test based on the measurement of serum CA 125 (a, log-scale), a logistic regression model LR2 (b) and Neural Network Model NN2 (c) (training + test data)

the data set[21]. The networks have a moderate complexity. The number of free variables in the model was kept small; the networks have only one hidden layer and only a few hidden neurons. Knowing that, in the present study, more than 30 different adnexal masses are encountered, we believe that, with more complex and enhanced architectures, a more detailed modelling of the probability of malignancy, and thus a better diagnostic performance, is possible. This would, however, require many more free variables and thus more advanced training techniques and a larger data set to limit the possibility of over-fitting.

We conclude that an ANN can be trained to predict malignancy from the patient's menopausal status and serum CA 125 levels, and some simple ultrasonographic indices of the adnexal mass. However, in order to test the generalizibility of the proposed neural network models, it is necessary to cross-validate them in a new and prospective trial before they can be recommended for use in routine clinical practice.

ACKNOWLEDGEMENTS

We are indebted to Professor Dr Ph. Moerman, Department of Pathology, University Hospitals Leuven, for expert help with histopathology. This research was partially carried out at the ESAT laboratory and the ICNN of the K.U. Leuven, in the following frameworks: the Program on Interuniversity Poles of Attraction (IUAP P4-02, P4-24) of the Belgian State, Prime Minister's Office – O.S.T.C.; a Concerted Action Project GOA-MIPS of the Flemish Community; the FWO (Fund for Scientific Research-Flanders) Research Communities ICCOS and Advanced Numerical Methods for Mathematical Modeling and the FWO project G.0262.97.

REFERENCES

1. Tingulstad S, Hagen B, Skjeldestad FE, Onsrud M, Kiserud T, Halvorsen T, Nustad K. Evaluation of a risk of malignancy index based on serum CA 125, ultrasound findings and menopausal status in the pre-operative diagnosis of pelvic masses. *Br J Obstet Gynaecol* 1996;103:826–31
2. Kehoe S, Powell J, Wilson S, Woodman C. The influence of the operating surgeon's specialisation on patient survival in ovarian carcinoma. *Br J Cancer* 1994;70:1014–17
3. Hacker NF, van der Burg MEL. Debulking and intervention surgery. *Ann Oncol* 1993;4(Suppl):17–22
4. Levin L, Lund B, Heintz APM. An overview of multi-variate analyses of prognostic variables with special reference to the role of cytoreductive surgery. *Ann Oncol* 1993;4(Suppl):23–9
5. Vergote IB, De Wever I. Primäre zytoreduktive Operation bei Ovarialkarzinom im fortgeschrittenen Stadium. Ist sie bei allen Patientinnen erforderlich? *Gynäkologe* 1997;30:102–7
6. Sassone AM, Timor-Tritsch IE, Artner A, Westhoff C, Warren WB. Transvaginal sonographic characterization of ovarian disease: evaluation of a new scoring system to predict ovarian malignancy. *Obstet Gynecol* 1991;78:70–6
7. Hata K, Hata T, Manabe A, Sugimura K, Kitao M. A critical evaluation of transvaginal Doppler studies, transvaginal sonography, magnetic resonance imaging, and CA 125 in detecting ovarian cancer. *Obstet Gynecol* 1992;80:922–6
8. Valentin L, Sladkevicius P, Marsál K. Limited contribution of Doppler velocimetry to the differential diagnosis of extrauterine pelvic tumors. *Obstet Gynecol* 1994;83:425–33
9. Tekay A, Jouppila P. Blood flow in benign ovarian tumors and normal ovaries during the follicular phase. *Obstet Gynecol* 1995;86:55–9

10. Timmerman D, Bourne TH, Tailor A, Collins WP, Verrelst H, Vandenberghe K, Vergote I. A comparison of methods for the pre-operative discrimination between benign and malignant adnexal masses: the development of a new logistic regression model. *Am J Obstet Gynecol* 1999; in press
11. Canis M, Mage G, Pouly JL, Wattiez A, Manhes H, Bruhat MA. Laparoscopic diagnosis of adnexal cystic masses: a 12 year experience with long-term follow-up. *Obstet Gynecol* 1994;83:707–12
12. Mais V, Ajossa S, Piras B, Marongiu D, Guerriero S, Melis BG. Treatment of nonendometriotic benign adnexal cysts: a randomized comparison of laparoscopy and laparotomy. *Obstet Gynecol* 1995;86:770–4
13. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol* 1990;97: 922–9
14. Granberg S, Wikland M, Jansson I. Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis: criteria to be used for ultrasound evaluation. *Gynecol Oncol* 1989;35:139–44
15. Tailor A, Jurkovic D, Bourne TH, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. *Ultrasound Obstet Gynecol* 1997;10:41–7
16. Haykin S. *Neural Networks: a Comprehensive Foundation*. New York: Macmillan Publishing Company, 1994
17. Shepherd GM, Koch C. Introduction to synaptic circuits. In Shepherd GM, ed. *The Synaptic Organization of the Brain*. New York: Oxford University Press, 1990:3–31
18. Newey VR. Classical versus artificial neural network analysis (Opinion). *Ultrasound Obstet Gynecol* 1997;10:5–8
19. Petersson F, ed. Annual report on the results of treatment in gynecological cancer. *Int J Gynecol Obstet* 1991. Twenty-first volume statements of results obtained in patients 1982 to 1986, inclusive 3- and 5-year survival up to 1990: S238–S277
20. Hanley JA, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36
21. Hanley JA, McNeil B. A method of comparing the areas under the receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–43