

Research Article

Comparison of Artificial Neural Network with Logistic Regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient Outcomes

Valérie Bourdès,^{1,2} Stéphane Bonnevey,³ Paolo Lisboa,⁴ Rémy Defrance,⁵ David Pérol,⁶ Sylvie Chabaud,⁶ Thomas Bachelot,⁶ Thérèse Gargi,⁶ and Sylvie Négrier⁶

¹ THEMIS, 60 avenue Rockefeller, 69 008 LYON, France

² ICTA-PM, 11 rue du Bocage, 21 121 Fontaine les Dijon, France

³ Laboratoire ERIC, Ecole Polytechnique Universitaire de Lyon 1, 69622 Villeurbanne, France

⁴ School of Computing and Mathematical Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

⁵ Oncology Department, Pfizer France, Avenue Dr. Lannelongue, 75014 Paris, France

⁶ Centre Léon Bérard, 28 rue Laennec, 69373 Lyon Cedex 08, France

Correspondence should be addressed to Valérie Bourdès, valerie.bourdes@icta.fr

Received 4 February 2010; Accepted 5 June 2010

Academic Editor: Tomasz G. Smolinski

Copyright © 2010 Valérie Bourdès et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The aim of this study was to compare multilayer perceptron neural networks (NNs) with standard logistic regression (LR) to identify key covariates impacting on mortality from cancer causes, disease-free survival (DFS), and disease recurrence using Area Under Receiver-Operating Characteristics (AUROC) in breast cancer patients. From 1996 to 2004, 2,535 patients diagnosed with primary breast cancer entered into the study at a single French centre, where they received standard treatment. For specific mortality as well as DFS analysis, the ROC curves were greater with the NN models compared to LR model with better sensitivity and specificity. Four predictive factors were retained by both approaches for mortality: clinical size stage, Scarff Bloom Richardson grade, number of invaded nodes, and progesterone receptor. The results enhanced the relevance of the use of NN models in predictive analysis in oncology, which appeared to be more accurate in prediction in this French breast cancer cohort.

1. Introduction

Artificial Neural Networks (ANNs) have been extensively used in many research areas from marketing to medicine [1]. They first received much attention from computer scientists, neurophysiologists, psychologists, and engineers, interested in biological nervous system organization and artificial intelligence. Their two main applications in medicine are pattern recognition (classification) and prediction: during these last years (from 1990s and increasing in the 2000s), the applications for prognostic and diagnostic classification in medicine have attracted growing interest in the medical literature. They have been applied to make predictions in numerous fields such as cardiology, molecular biology, trauma

outcomes, neonatology, and oncology (acute myeloma, prostatic cancer, colon cancer, and breast cancer) [2–7]. A review of evidence of ANN benefit in the medical field has been published [1].

ANNs are particularly useful in prediction where highly nonlinear approaches are required to sift through the plethora of available information. They present the main advantage of not being based on “*a priori*” assumptions and of allowing detection of links between factors that conventional statistical techniques such as logistic regression may not be able to detect. With the increasing number of potential prognostic factors for breast cancer, it is becoming increasingly more difficult to integrate the combination of these factors into an accurate prediction of individual clinical

course. The main pragmatic impact of allocating patients into prognostic risks is the disease management with the choice of treatment.

The logistic regression in this work was chosen as an accepted standard for prediction by biostatisticians [8] in order to evaluate the neural network.

2. Materials and Methods

2.1. Data Base Recruitment and Followup. Since 1996, all patients whose initial surgical treatment (lumpectomy or mastectomy) was performed at the Centre Léon Bérard (CLB) (primary or secondary following neoadjuvant chemotherapy), based in Lyon, have been registered in the database. Tumour diagnosis was confirmed by histology and concerned infiltrating or *in situ* carcinoma; the data were collected by a clinical research assistant (CRA) at the CLB from the patient's computer medical record file. Followup was provided until patient's death (letter to the referring physician, or registrar's office) as well as information on the evolution of the tumour in terms of local or distal recurrences. An in-house algorithm has been developed to detect the administrative area where the patients live. This breast cancer database was updated regularly according to the consecutive clinical cases treated at the CLB. About 220 explanatory variables have been captured, including the clinical or surgical history of the patient, the histology of the tumour, the treatments applied, and some immunohistochemical covariates (hormonal status, Her2+). By the end of March 2006, a total of 4,070 events were stored in the database, corresponding to 3,929 patients.

2.2. Cohort Selection. In order to work on a more homogeneous type of malignant disease, records concerning *in situ* carcinomas without infiltrating component were not taken into account in this analysis. As some patients had more than one record (141), only the first episode defined by the earlier date of diagnosis was selected. For 40 patients, two records were available for the first diagnosis, corresponding to bilateral tumours. These cases were not included into the analysis as we could not decide which histology was more influential on survival parameters. With the same objective, namely, to compare only "pure" cases, patients with a history of ipsilateral or contralateral carcinoma not treated at the CLB were not taken into account. As the aim of this project was to work out new prognosis tools, it was decided to leave out patients with initial metastatic carcinoma who were considered as specific and with a poor prognosis. Finally, all the patients with a date of diagnosis prior to December 31st, 2004 were included. After this cohort selection, 2,535 records corresponding to 2,535 different patients were selected for this work. A total of 32 parameters including clinical, histological, immunohistochemical, and treatment variables were considered as relevant by the clinicians and were used for the analysis (Table 1).

2.3. Events of Interest to Be Analysed. The following four events of interest were defined and analysed: mortality

TABLE 1: Variables included in the database.

Variables	Modalities
Breast side	Right, Left
Age	<i>Real</i>
Delay diagnosis—1st treatment	<i>Time</i>
Personal cancer history	No, Yes
Menopausal status	No, Yes
Cutaneous inflammation	No, Yes
Skin invasion	No, Yes
Clinical number of nodules	0, 1, ≥ 2
Clinical T stage (size)	T0, T1, T2, T3, T4
Clinical N stage	N0, N+
Number of tumours (histology)	0 or 1, ≥ 2
Tumour size (mm) (histology)	<i>Real</i>
Histological type	Lobular, Ductal, Mixed, Microinvasive, Others
SBR grade	1, 2, 3
Lymphatic embolus	No, Yes
Nervous colonisation	No, Yes
Necrosis (histology)	No <i>in situ</i> component, <i>In situ</i> component without necrosis <i>In situ</i> component with necrosis
Infiltrating tumour associated with CIS	Others, >75% <i>in situ</i>
Limits of exeresis—infiltrating carcinoma	Not <i>in sano</i> , <i>In sano</i>
Skin—infiltrating tumour	No or Not Applicable, Yes
Skin—embolus	No or Not Applicable, Yes
Nipple—infiltrating tumour	No or Not Applicable, Yes
Nipple—cancer in situ	No or Not Applicable, Yes
Nipple—embolus	No or Not Applicable, Yes
Nipple—Paget's disease	No or Not Applicable, Yes
Pectoral muscle invaded	No or Not Applicable, Yes
Number of invaded nodes	0, 1 or 2, ≥ 3
Nodes invaded	N0, Micro metastasis, Macrometastasis
Capsular breaking	No, Yes
Oestrogen receptor: % marked cells (count)	<10%, (10–50)%, >50%
Oestrogen receptor (intensity)	+, ++/+++
Progesterone receptor: % marked cells (count)	<10%, [10–50]%, >50%
Progesterone receptor (intensity)	+, ++/+++

attributed to cancer causes (specific mortality) (136 patients, 5.4%), disease-free survival (DFS) (372 patients, 14.8%), local recurrence (113 patients, 4.5%), and metastatic distal recurrence (242 patients, 9.6%).

For all analyses, a patient was considered to have a recurrence if the patient's status in the followup form was not ticked as "complete response", if an "evolution form" for this patient was filled up, or if there was a second record in the initial breast database for this patient. Local recurrence was confirmed if an "evolution form" with the local part was fulfilled or if there was a second record in the initial breast database for this patient, with a clinical stage assessed as M0. Distal recurrence was confirmed if an "evolution form" with the distal part was fulfilled or if there was a second record in the initial breast database for this patient, with clinical stage assessed as M+.

2.4. Variables Selection of Neural Networks. To compare logistic regression (LR) and neural networks (NNs) models, many papers use the same variables for both input models (the variables selected by the multivariate analysis) [9]. This choice is justified by the large degree of overlap between the sets of variables selected with both approaches. But in this paper, we decided to build two NN models for each analysis according to their inputs: the first one with the multivariate selected variables (used for the LR models, i.e., NN-varLR) and the second one with the variables selected with our NN approach, i.e., NN-varNN). To select the most significant variables for use with neural networks, we used three different methods: forward and backward stepwise feature selection and a genetic input selection algorithm. Forward selection consists in choosing the most predictable variable then checks for a second variable that, added to the first, most improves the model; this process is repeated until either all variables have been selected or no further improvement is made. Backward stepwise feature selection is the reverse process: it starts with all the variables and then removes a variable at each stage which less degrades the model. Genetic algorithm selection is a heuristic seeking the optimal set of input variables. This heuristic builds a model by a succession of artificial transformations (mutation, crossover, and selection) from an initial population of variables sets. Each of our genetic selections was made from a population of 100 individuals (one individual corresponds to one set of variables) on 100 generations. Each set of variables corresponds to a binary string where a 0 indicates that the variable is not in the set of variables, and a 1 indicates that the variable is in this set. This set is tested with the help of a neural network, and the objective function is the error of this neural network on a training set.

Each of these methods has some advantages and disadvantages. Forward selection is faster than the others, but it may miss key variables if they are interdependent. Backward selection does not suffer from this problem, but it is time consuming at the beginning of the process due to the evaluations of the whole set of variables. Genetic algorithm selection is the slowest method.

With our choice (100 individuals and 100 generations), it performs 10,000 evaluations of sets of variables. For example, the selection of one set of variables is about 60 times longer with the genetic algorithm than the backward selection, but genetic algorithms are well suited for feature selection as there is a large number of possible variables. Because of their differences and complementarities, we decided to combine these three methods to select the inputs of our NN models.

With a view to improving the generalization capability of networks and to decreasing the network size and execution size, a penalty can be used to penalise the large sets of variables [10]. In this way, a penalty parameter is multiplied by the number of selected variables and added to the error level. Following different analyses, we finally used a small penalty equal to 0.0001 with half of selection algorithms and no penalty with the other half selection algorithms. Each method calculates the value of the set of covariates selected at each step while building the neural network. As, during the learning and test phases, data are chosen and presented at random, then, training and validation sets are different, and then results are different too. According to some experiments, we have chosen to perform each method 40 times: 20 without penalty and 20 with a penalty equal to 0.0001. Indeed, 120 (3×40) selections were performed. In order to illustrate our variables selections, Table 2 shows the selection results for the prediction of "mortality from cancer causes".

For each method, Table 2 indicates how many times these variables have been selected and the corresponding percentage. The last column "Global" gives the total number of selections. For example, "Nervous Spread" has never been selected by the forward selection method with a 0.0001 penalty. It has been selected 11 times by the same method without penalty. The last row shows that the "Progesterone Receptors" variable has always been selected by all methods. The last column of this table takes into account results selection combination of all the methods. The problem is to determine a threshold α that decides whether a variable will be included or not in our model. To fix this threshold, we built and evaluated some NN models with different combinations of variables according to their global percentage. According to these experimental results, we decided that a variable would be kept as an input for our model if it was selected in at least 95% of our 120 selections. Moreover, this value of 95% leads to the inclusion of a reasonable number of variables regarding the complexity of the NN models. This final selection obtained was validated and approved by the oncologists of Centre Leon Berard.

2.5. Building Neural Networks. ANNs have been applied in a wide range of problems and have given, in many cases, superior results to standard statistical models [11]. In particular, the predictive reliability of ANN models has been demonstrated in medical diagnosis [12]. According to the literature and some previous experimental analyses, we decided to use the Multilayer Perceptrons (MLP) for predictions [13]. In this work, we used only one type of ANNs for different reasons. First our study was dealing with 4 different analyses, and, for each one, two NN models were

TABLE 2: Details of variables selections for “specific mortality”.

	Forward		Forward		Backward		Backward		Genetic		Genetic		Global	
	0.0001		0.0		0.0001		0.0		0.0001		0.0		#	%
	#	%	#	%	#	%	#	%	#	%	#	%		
Side of Breast	0	0%	4	20%	0	0%	4	20%	0	0%	4	20%	12	10%
History of Cancer	0	0%	11	55%	0	0%	10	50%	1	5%	4	20%	26	22%
Menopausal	16	80%	19	95%	19	95%	17	85%	17	85%	19	95%	107	89%
Cutaneous inflam	0	0%	4	20%	0	0%	3	15%	0	0%	3	15%	10	8%
Skin invasion	6	30%	14	70%	3	15%	11	55%	5	25%	13	65%	52	43%
Clin. Nb. Nodules	20	100%	20	100%	20	100%	20	100%	20	100%	20	100%	120	100%
Invaded Nodes	20	100%	19	95%	20	100%	19	95%	20	100%	20	100%	118	98%
Stade N	16	80%	19	95%	18	90%	18	90%	14	70%	20	100%	105	88%
Nb of tum (histo)	8	40%	16	80%	6	30%	20	100%	5	25%	18	90%	73	61%
Size of tum (histo)	0	0%	4	20%	0	0%	5	25%	0	0%	3	15%	12	10%
SBR Grade	19	95%	20	100%	19	95%	20	100%	20	100%	20	100%	118	98%
Lymphatic Embolus	14	70%	15	75%	13	65%	18	90%	13	65%	14	70%	87	73%
Nervous spread	0	0%	11	55%	4	20%	7	35%	1	5%	6	30%	29	24%
Limits Exeresis	0	0%	10	50%	0	0%	8	40%	0	0%	5	25%	23	19%
Skin infiltrating	2	10%	9	45%	7	35%	8	40%	0	0%	11	55%	37	31%
Skin embolus	0	0%	6	30%	0	0%	2	10%	0	0%	2	10%	10	8%
Nipple infiltrating	1	5%	4	20%	0	0%	4	20%	0	0%	3	15%	12	10%
Nipple Cancer in situ	0	0%	4	20%	0	0%	1	5%	0	0%	1	5%	6	5%
Nipple embolus	2	10%	5	25%	0	0%	3	15%	1	5%	3	15%	14	12%
Nipple Paget	0	0%	9	45%	0	0%	10	50%	0	0%	5	25%	24	20%
Pectoral muscle inv	0	0%	5	25%	0	0%	3	15%	0	0%	3	15%	11	9%
Age	6	30%	18	90%	5	25%	19	95%	7	35%	19	95%	74	62%
Time diag-firstTreat	0	0%	2	10%	0	0%	2	10%	0	0%	0	0%	4	3%
Time diag-firstSurgery	0	0%	8	40%	0	0%	11	55%	0	0%	6	30%	25	21%
Histology	19	95%	19	95%	20	100%	20	100%	20	100%	20	100%	118	98%
Necrosis	11	55%	16	80%	13	65%	16	80%	15	75%	13	65%	84	70%
Infiltrating tumour	0	0%	8	40%	0	0%	11	55%	1	5%	11	55%	31	26%
Clinical Size	20	100%	20	100%	19	95%	20	100%	20	100%	19	95%	118	98%
Capsular breaking	18	90%	19	95%	12	60%	19	95%	17	85%	18	90%	103	86%
Oestro receptors	19	95%	18	90%	17	85%	20	100%	17	85%	16	80%	107	89%
Progest receptors	20	100%	20	100%	20	100%	20	100%	20	100%	20	100%	120	100%
Nb. Nodes Invad	20	100%	20	100%	20	100%	20	100%	20	100%	20	100%	120	100%

built for both variables selections; secondly, MLP is the most commonly used ANN. For each analysis and each set of input variables, we built a three-layer network with an input layer corresponding to our risk factors (selected variables), a hidden layer with hyperbolic activation functions, and one linear output unit modelling the dichotomous risk outcome (Figure 1).

The number of neurons on the hidden layer was determined according to the number and the nature of variables at the entry. Weights and bias of neural network are determined by training with a two-phase procedure. The first phase is a quite short burst of backpropagation, with a moderate training rate. The second phase is a longer run of conjugated gradient descent, a much more powerful algorithm, which is less likely to encounter convergence problems than otherwise due to the use of backpropagation first. During this learning process, the weights in a MLP are adjusted using least squares

fitting together with the training two-phase procedure to minimize a root mean square error function. In order to interpret the network outputs as probabilities and to make them comparable to the results of logistic regression, we used a cross entropy error function to adjust weights. This cross entropy function is specially designed for classification problems where it is used in combination with hyperbolic activation function [14].

A continuous input value is prescaled to a range between 0 and 1; a two-state nominal variable, which corresponds to one entry of the neural network, is represented by transformation into a numeric value (e.g., “Skin invasion” = 0 or 1); a many-state nominal variable is recoded into as many binary entries as modalities (on Figure 1, x_1 is a 5-state nominal variable which correspond to 5 inputs, and x_n is a 3-state nominal variable which correspond to 3 inputs).

TABLE 3: Main characteristics of the cohort.

Clinical parameters		
Age		
Mean \pm SD		59.8 \pm 12.1
Median [range]	<i>N</i> = 2535	54.0 [23.2; 91.7]
Breast side		
Left	<i>N</i> = 2535	50.1%
Right		49.9%
Personal cancer history*	<i>N</i> = 2535	2.7%
Menopausal status	<i>N</i> = 2506	68.7%
Skin invasion	<i>N</i> = 2531	4.4%
Cutaneous inflammation	<i>N</i> = 2529	1.9%
Clinical number of nodules		
0		22.9%
1	<i>N</i> = 2525	74.3%
≥ 2		2.7%
Clinical T stage (size)		
T0		22.2%
T1		44.9%
T2	<i>N</i> = 2573	22.0%
T3		4.9%
T4		5.9%
Clinical N stage		
N+	<i>N</i> = 2525	17.3%
Histological parameters		
Number of tumours (histology)		
0 or 1	<i>N</i> = 2459	88.2%
≥ 2		11.8%
Tumour size (mm) (histology)		
Mean \pm SD	<i>N</i> = 2452	24.1 \pm 20.1
Median (range)		20.0 [0.4; 250.0]
Histological type		
Lobular		10.0%
Ductal		73.6%
Mixed	<i>N</i> = 2535	3.7%
Micro invasive		1.5%
Others		11.2%
Histological grade (SBR)		
1		23.5%
2	<i>N</i> = 2449	47.2%
3		29.3%
Lymphatic embolus	<i>N</i> = 2523	36.8%
Nervous colonisation	<i>N</i> = 2520	13.7%
Necrosis (histology)		
No <i>in situ</i> component		19.5%
<i>In situ</i> component without necrosis	<i>N</i> = 2534	35.0%
<i>In situ</i> component with necrosis		45.4%
Infiltrating tumour associated with CIS		
>75% <i>in situ</i>	<i>N</i> = 2535	5.5%
Limits of exeresis—infiltrating carcinoma		
Not <i>in sano</i>	<i>N</i> = 2496	2.5%

TABLE 3: Continued.

Skin—infiltrating tumour	<i>N</i> = 2535	5.6%
Skin—embolus	<i>N</i> = 2535	2.0%
Nipple—infiltrating tumour	<i>N</i> = 2535	7.7%
Nipple—cancer <i>in situ</i>	<i>N</i> = 2535	6.8%
Nipple—embolus	<i>N</i> = 2535	4.0%
Nipple—Paget's disease	<i>N</i> = 2535	1.3%
Pectoral muscle invaded	<i>N</i> = 2535	1.4%
Number of invaded nodes:		
0		52.0%
1 or 2	<i>N</i> = 2515	28.2%
≥ 3		19.8%
Invaded nodes:		
N0		8.5%
Micrometastasis	<i>N</i> = 2535	58.9%
Macrometastasis		32.7%
Capsular breaking	<i>N</i> = 2535	26.0%
Immunohistochemical parameters		
Oestrogen receptor: % marked cells (count):		
<10%		17.1%
10–50%	<i>N</i> = 2443	7.2%
$\geq 50\%$		75.8%
Oestrogen receptor (intensity):		
+	<i>N</i> = 2030	25.8%
++ and +++		74.2%
Progesterone receptor: % marked cells (count)		
<10%		28.2%
10%–50%	<i>N</i> = 2443	17.4%
$\geq 50\%$		54.4%
Progesterone receptor (intensity):		
+	<i>N</i> = 1796	17.4%
++ and +++		82.6%

*69 patients had cancer history. Localizations were 16 gynaecologic, 11 digestive, 11 Hodgkin/LMNH, 8 melanoma, 7 thyroid, 4 lung, 4 head and neck, 4 urinary, and 3 others.

The objective is to generate neural networks not too close to the data used for the learning phase (to avoid overfitting) in order to build a consistent predictor that can be used with other data (those not used for the learning phase). To obtain networks with a strong capacity of generalisation, we divided data randomly in two datasets:

- (i) a learning set to build the models (LR and NN),
- (ii) a testing set for the evaluation (this set is not used for construction).

The learning set was composed of 1,775 individuals (about 70% of total population), with 153 deaths (72.2% of death) and 267 disease-free survivals (71.2% of disease-free survival). The testing set was composed of 760 individuals,

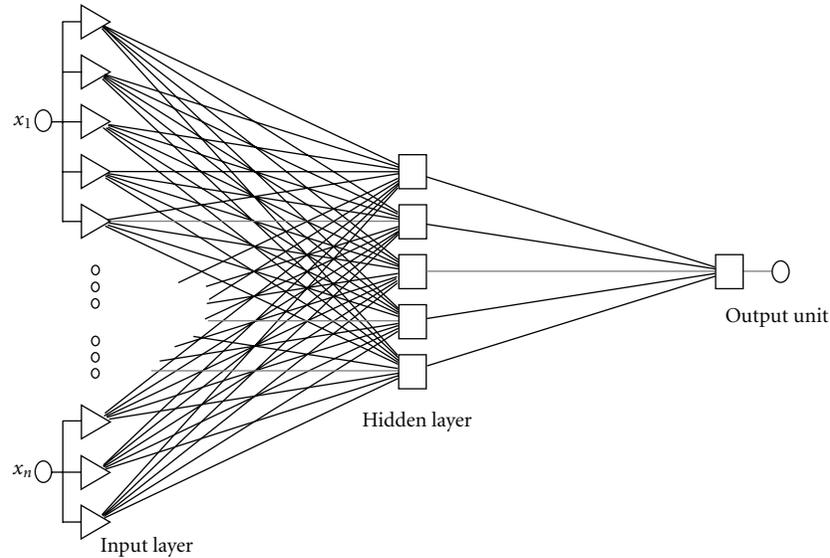


FIGURE 1: Architecture of a feed-forward supervised 3-Layer Perceptron neural network.

59 deaths, and 105 disease-free survivals. For this paper, one LR model and two NN models were built for each event to analyse. The first neural network model was built according to the logistic regression inputs (selected variables from logistic regression analysis), and the second one with the neural network selected variables approaches.

2.6. Statistics. With a view to selecting the prognostic factors of the LR model, an univariate logistic analysis was first performed for each event. Then, all the variables significant at the level of 10% were included in the multivariate logistic step. Neural network constructions, as well as the selection of significant covariates for these models, were performed with the Statistical Neural Networks software release 7.1. The logistic regression was performed with SAS Software 9.1. A total of 36 variables were extracted from the database: 32 covariates (Table 1) and 4 events of interest. To prevent introducing bias we decided that not all the surgical and treatment variables were taken into account for both LR and NN analysis. A total of 2,535 observations were analysed. Only 0.6% missing data were recorded in the CLB database. Due to their paucity, they were not coded as a separate attribute, and only the available data were used for modelling.

The comparison between the three models, LR, NN with inputs of LR selection (NN-varLR), and NN with inputs of NN selection (NN-varNN), was assessed using AUROC (Area Under Receiver-Operating Characteristics). The area under curve is a good measure of the overall predictive accuracy of an analytic tool. It represents a plot of sensitivity *versus* 1 minus specificity. Sensitivity measures the fraction of positive cases classified as positive, and specificity measures the fraction of negative cases classified as negative. ROC indicator in this work measures the separation between the probability distributions of the output neuron activation under the null hypothesis (no event at the end of October

2004) and under the alternative hypothesis (event at the end of December 2004).

3. Results

3.1. Cohort Analysis. The main characteristics of the cohort are described in Table 3. Median age was 54 years ranging from 23 to 92 years. A total of 69 patients (2.7%) had a history of previous cancer; 69% of patients were in menopausal status. Clinical T stage was >2 in 10% of the cases, and clinical N stage was positive in 17% of patients. In 12% of the cases the diagnosis was multi focal. The histological median tumour size was 20 mm with a range from 0.4 to 250 mm. The carcinoma was found ductal in more than 70% of the cases with a SBR (Scarff Bloom Richardson) grade 1, 2, and 3 in 24%, 47%, and 30%, respectively. The percentage of marked cells was \geq to 50% for progesterone receptors (PRs) in 54% of the cases and for oestrogen receptors (ERs) in 76%.

Overall mortality occurred in 8.4% of cases and specific mortality in 5.4% of cases. A total of 316 progressions were notified (12.5%) representing at least 4.5% of local recurrences and 8.4% of metastatic events. The median followup was 4.1 years (CI95% = 4.0–4.5) with a maximum followup of 10.2 years.

3.2. Events. Table 4(b) displays the results of the selections from statistical and neural networks approaches for “Specific Mortality”. A variable was chosen by the neural approach if the percentage of selection was greater or equal to 95% (bold data of column NN), and a variable was chosen by the logistic regression analysis when a cross is matched in the corresponding column LR. Four variables among the five selected by the logistic regression were retained by both approaches, either LR or the NN approach, namely,

TABLE 4: Variables selections for Logistic Regression and Neural Network approaches.

(a) Disease-free survival.		
Variables	NN	LR
Invaded nodes	100%	X
Clinical size stage	100%	X
Number nodes invaded	100%	X
SBR grade	98%	X
Histology	98%	
Necrosis	98%	
Oestrogen receptors	74%	X
Skin infiltrating tumour	21%	X

(b) Mortality from cancer causes.		
Variables	NN	LR
Clinical number nodules	100%	
Progesterone receptor	100%	X
Number nodes invaded	100%	X
Clinical size stage	98%	X
SBR grade	98%	X
Histology	98%	
Invaded nodes	98%	
Skin embolus	8%	X

(c) Local recurrence.		
Variables	NN	LR
Number nodes invaded	98%	X
Lymphatic embolus	95%	X
Ganglion invaded	95%	
Necrosis	95%	X
Oestrogen receptors	95%	X
Histology	48%	X
Number tumours	15%	X
Skin invasion	13%	X

(d) Metastatic recurrence.		
Variables	NN	LR
Clinical Size stage	100%	X
Invaded nodes	100%	
Progesterone receptors	100%	X
Number of nodes invaded	100%	X
SBR grade	98%	X
Histology	98%	
Oestrogen receptors	85%	X

“Progesterone receptor”, “Number nodes invaded”, “Clinical size stage”, and “SBR grade”. In addition to these common variables, the NN approach selected, with high percentages, the variables “Histology”, “Invaded nodes”, and “Clinical number of nodules”. On the other hand, the multivariate analysis selected the “Skin embolus” whereas this variable was only selected in 8% of cases by the NN selection.

TABLE 5: Sensitivity and specificity.

(a) Disease-free survival.			
	LR	NN-varLR	NN-varNN
Sensitivity	71.5	82.2	80.3
Specificity	68.9	77.9	79.3
False negative rate	28.5	17.8	19.7
False positive rate	31.1	22.1	20.7
Positive predictive value	11.9	18.0	18.6

(b) Mortality from cancer causes.			
	LR	NN-varLR	NN-varNN
Sensitivity	80.5	86.7	87.5
Specificity	77.9	76.0	77.3
False negative rate	19.5	13.3	12.5
False positive rate	22.1	24.0	22.7
Positive predictive value	17.4	17.2	18.2

(c) Local recurrence.			
	LR	NN-varLR	NN-varNN
Sensitivity	68.6	72.4	72.9
Specificity	65.1	66.7	64.9
False negative rate	31.4	27.6	27.1
False positive rate	34.9	33.3	35.1
Positive predictive value	8.3	9.1	8.8

(d) Metastatic recurrence.			
	LR	NN-varLR	NN-varNN
Sensitivity	71.3	70.6	71.9
Specificity	78.1	78.8	78.5
False negative rate	28.7	29.4	28.1
False positive rate	21.9	21.2	21.5
Positive predictive value	25.9	26.4	26.4

Tables 4(a), 4(c), and 4(d) display variables selections from LR and NN approaches of “Disease free survival”, “Local recurrences”, and “Distal recurrences” analyses.

The architecture of MLP neural models for “Specific Mortality” analysis according to variable selection is the following.

(i) NN-varLR. The best model of NN obtained with the inputs of the LR selection (Table 4(b)) is described as follows.

- (a) One input layer corresponds to the 5 covariates. The skin embolus covariate, with 2 modalities, corresponds to a binary entry. The other covariates are recoded into as many entries as modalities (3 covariates with 3 modalities and 1 covariate with 5 modalities). The network built that way has 15 binary entries.
- (b) One hidden layer composed of 6 neurons with hyperbolic activation function.
- (c) One output layer composed of one neurone with logistic activation function.

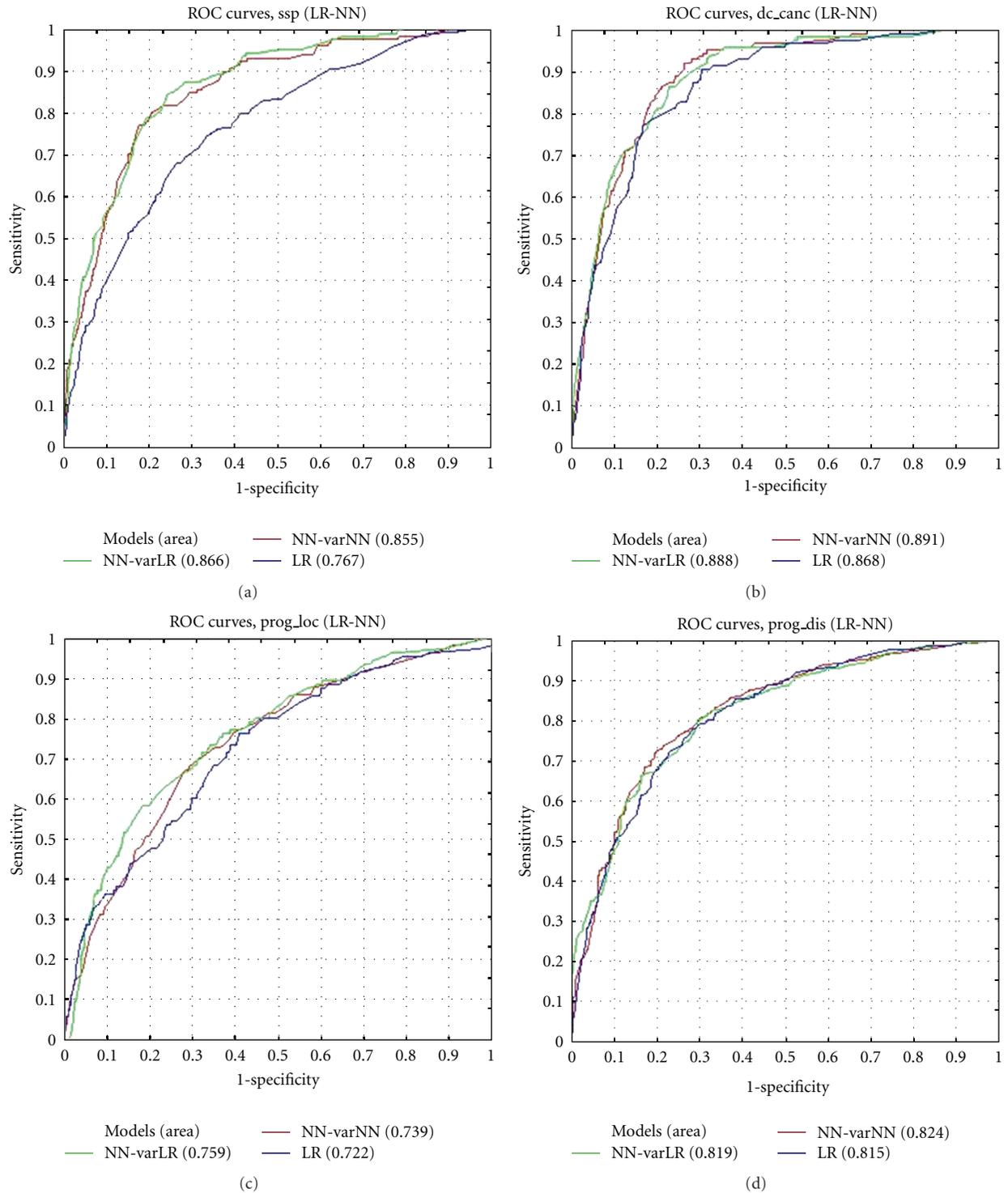


FIGURE 2: ROC curves from LR, NN-varLR, and NN-varNN models. (a) Disease-free survival. (b) Mortality from cancer causes. (c) Local recurrence. (d) Metastatic recurrence.

The same types of MLP were built for “Disease free survival”, “Local recurrences”, and “Distal recurrences” analysis.

- (ii) NN-varNN. The best model of NN obtained with the inputs of the NN selection (Table 4(b)) is described as follows.

- (a) One input layer corresponding to the 7 covariates. These covariates are recoded into as many entries as modalities (5 covariates with 3 modalities, 1 covariate with 4 modalities, and 1 covariate with 5 modalities). The network built that way has 24 binary entries.

- (b) One hidden layer composed of 6 hidden units with hyperbolic activation functions.
- (c) One output layer composed of one neurone with logistic activation functions.

The same types of MLP were built for “Disease free survival”, “Local recurrences”, and “Distal recurrences” analyses.

ROC Curves. According to “Specific Mortality”, the AUROC curves of “Specific Mortality” were very similar between the three models with a slight superiority in favour of the NN models (Figure 2(b)). The corresponding sensitivity and specificity optimal values are given in Table 5(b).

Figures 2(a), 2(c), and 2(d) display AUROC curves, and sensitivity and Tables 5(a), 5(c), and 5(d) display specificity optimal values for the event “Disease-Free survival”, “Local recurrences”, and “Distal recurrences” analyses.

4. Discussion

In order to best assess the comparison between LR and MLP predictions, we needed to gather several conditions regarding the cancer to study and the dataset to analyse. We needed for the cancer a specific combination of a well-described clinical course in the literature and a complex interaction between the covariates to introduce in the models. Breast cancer appeared to be the best one to fully achieve these conditions, and some authors have already shown NN could predict the probable clinical course of breast cancer patients [15, 16]. Regarding the dataset, the best data quality was required (to minimize missing and inconsistent data for the training of the NN), as well as a sufficient followup of the cohort. The median followup of 4.2 years appears reasonable enough even though it is slightly short. The database of the Leon Berard Centre appeared to be highly qualified because of a very small percentage of missing data, an inclusion criterion that is well defined, and a regular update of this database done by a dedicated CRA.

The consistency across the different models may be explained by the good quality dataset of the CLB database and emphasizes the relevance of the use of the ANN in predictive analysis in oncology.

Regarding the cohort selection, the patients retained for analysis were suffering from primary breast cancer and locally advanced cancer without metastasis. The majority of the cases (89%) were $\leq T2$ stage, and the histology was ductal carcinoma for 74% of them. The idea was to obtain a homogeneous cohort and to be in the situation to enable us to potentially identify prognostic factors. This situation excluded the very poor prognosis and explains the paucity of events to analyse (5.4% specific mortality and 12.5% for the total progressions including local and distal recurrences).

If we look at the clinical outputs of our variable selection, the four predictive factors commonly selected for the specific mortality analysis, by both LR and NN approaches, were the following: Clinical size stage, SBR grade, Number of nodes invaded, and Progesterone receptors. It is to be underlined that the first three variables are well known within the

medical literature and are related directly to clinical indicators routinely used by the NPI (Nottingham Prognostic Index). The NN models, either NN-varNN or NN-varLR, selected three additional factors, namely, Histology, Number of tumour nodules, and Invaded nodes (axillary lymph nodes). These results are compatible with published ones on other cohorts using Bayesian neural networks [7]. The other predictive factors we found with our NN selection other than those used in the NPI are the hormonal factors (PR and ER). Their role must be underlined here as the PR appeared to be a major predictive factor for the specific mortality, as well as for the distal recurrence study. In addition the ER appeared to be a major predictive factor for DFS and the local recurrence. The protective role of both receptors was already known even in terms of time-depending joint effects with tumour size or histology [2] but the respective role of PR and ER split into local and distal recurrences was not described so far. These results are worth exploring with more accuracy in further studies. Regarding the comparison between the performances of the three models, the main results we found consisted in showing that for breast cancer-specific mortality and DFS analysis, the areas under curves appear to be greater with the NN models than the LR (better sensitivity as well as specificity). These results underlined the predictive accuracy of the NN models comparing with the LR and, their relevance as predictive tools. Regarding the recurrence studies, the ROC curves were not as good as those of specific mortality and DFS. Nevertheless the AUROCs were somewhat similar between the LR and NN models, with a slight improvement in favour of the NN models.

As previously described, we decided to choose a MLP Neural Network for this specific work. One reason is that MLP is the ANN most widely applied to real-world problems in medical diagnosis and prediction; because of the numerous NN models built in this study, only one type of ANN is used. Other types of NN may have been used such as the Probabilistic Neural Network (PNN). PNN is particularly adapted to stepwise procedures aimed at selecting and classifying prognostic factors from small datasets [17]. Moreover some authors consider MLP to be superior to PNN [9].

One of the criticisms towards neural networks is that their process inside is unknown, and some authors consider them as “black boxes”. In order to prevent this criticism and also to enhance as much as possible the NN selection process we decided to use several variables selection techniques perfectly coded and to use one penalty for some selection methods.

Using these three different techniques for the variables selection may be criticized as a time-consuming process. It was here a guarantee for the best selection and for avoiding overfitting, which is the main limit of the neural networks. The penalty was used to penalise the big sets of variables with a view to improving the generalization capability of networks. The value of the penalty chosen had a substantial impact on the variable selection. We decided therefore to perform each method half of the time without any penalty and the other half of the time with a penalty equal to 0.0001.

A total of 120 selections per covariate were performed with each method of variable selection. This total number of selections increases considerably the time variable selection process but we expected a gain in improvement of the selection. The results we obtained tended to confirm our choice by showing greater AUROC for the mortality and DFS analysis particularly.

A limitation of our work is that there is a need for an external validation with a second and independent dataset. We are planning to carry out this extension of the project, and for this reason we think our results should be considered as exploratory rather than predictive. The main objective was to compare standard predictive tools to innovative ones in the medical field, and in oncology in particular. This work brought some clinical insights to be confirmed further in the field of breast cancer prognosis. Another limitation, and probably the major one, is that the present study did not investigate complex time-depending effects of prognostic factors of breast cancer over followup time. The definition of risk categories based on tumours or patients' characteristics may evolve in the course of the followup according to the disease dynamics. Some authors developed the PLANN (with a partial logistic artificial neural network) approach for the analysis of the hazard function as a function of time and covariates for censored survival time data [2] showed that patients with small tumours with high ER levels and Invasive Ductal Carcinoma plus Invasive Lobular Carcinoma histology could be at high risk of disease recurrence in the medium to long term and consequently should be carefully monitored. They also showed a joint time-dependent effect of histology and ER. Additional analysis taking into account all censored variables must be carried out to improve these models predictivity.

In conclusion, this paper presents an evaluation tool for the prognosis of breast cancer on a cohort of nonmetastatic patients using clinical, pathological, and immunohistochemical data. The results of this work, whose main aim was to compare the LR and NN performances in predictions, have to be considered as exploratory results rather than conclusive results for predictions. Our neural network selection approach highlights some different inputs for the models from classical statistical selections. All our input selections were validated by clinicians. We hope this work will convince clinicians to use commonly ANN for the extraction of large dataset patterns in prognostic factors, at risk group definitions, and to envisage these tools as a decision support to appropriate treatments for the individual patient. ANN should be considered powerful predictive tools, to be routinely added to standard logistic regression. The next step will be the development of a web-based tool for community use.

Acknowledgments

This paper was financially supported by Pfizer France and was carried out as a collaboration between THEMIS-ICTA Group, Centre Léon Bérard, Liverpool John Moores University, and LIRIS- Lyon 1 University.

References

- [1] P. J. G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention," *Neural Networks*, vol. 15, no. 1, pp. 11–39, 2002.
- [2] E. Biganzoli, P. Boracchi, D. Coradini, M. G. Daidone, and E. Marubini, "Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors," *Annals of Oncology*, vol. 14, no. 10, pp. 1484–1493, 2003.
- [3] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in Medicine*, vol. 17, no. 10, pp. 1169–1186, 1998.
- [4] E. Biganzoli, P. Boracchi, and E. Marubini, "A general framework for neural network models on censored survival data," *Neural Networks*, vol. 15, no. 2, pp. 209–218, 2002.
- [5] S. D. Dolgobrodov, R. Marshall, P. Moore, R. Bittern, R. J. C. Steele, and A. Cuschieri, "e-science and artificial neural networks in cancer management," *Concurrency and Computation: Practice and Experience*, vol. 19, no. 2, pp. 251–263, 2007.
- [6] H. B. Burke, P. H. Goodman, D. B. Rosen et al., "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer*, vol. 79, no. 4, pp. 857–862, 1997.
- [7] P. J. G. Lisboa, H. Wong, P. Harris, and R. Swindell, "A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer," *Artificial Intelligence in Medicine*, vol. 28, no. 1, pp. 1–25, 2003.
- [8] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2000.
- [9] R. Voss, P. Cullen, H. Schulte, and G. Assmann, "Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks," *International Journal of Epidemiology*, vol. 31, no. 6, pp. 1253–1262, 2002.
- [10] H. Zeng and H. J. Trussell, "Feature selection using a mixed-norm penalty function," in *Proceedings of the International Conference on Image Processing (ICIP '06)*, October 2006.
- [11] H. White, "Learning in artificial neural networks: a statistical approach," *Neural Computation*, vol. 1, pp. 425–464, 1989.
- [12] W. G. Baxt, "Application of artificial neural networks to clinical medicine," *Lancet*, vol. 346, no. 8983, pp. 1135–1138, 1995.
- [13] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [14] D. M. Kline and V. L. Berardi, "Revisiting squared-error and cross-entropy functions for training neural network classifiers," *Neural Computing and Applications*, vol. 14, no. 4, pp. 310–318, 2005.
- [15] P. M. Ravdin and G. M. Clark, "A practical application of neural network analysis for predicting outcome of individual breast cancer patients," *Breast Cancer Research and Treatment*, vol. 22, no. 3, pp. 285–293, 1992.
- [16] P. M. Ravdin, G. M. Clark, S. G. Hilsenbeck et al., "A demonstration that breast cancer recurrence can be predicted by Neural Network Analysis," *Breast Cancer Research and Treatment*, vol. 21, no. 1, pp. 47–53, 1992.
- [17] J. M. Le Goff, L. Lavayssière, J. Rouëssé, and F. Spyrtos, "Nonlinear discriminant analysis and prognostic factor classification in node-negative primary breast cancer using probabilistic neural networks," *Anticancer Research*, vol. 20, no. 3, pp. 2213–2218, 2000.