Validation procedures in radiological diagnostic models. Neural network and logistic regression.[1]

Estanislao Arana[2], Pedro Delicado[3] and Luis Martí-Bonmatí[4]

[1] To appear in "Investigative Radiology", October 1999.

[2] Department of Radiology. Hospital Casa de Salud. Valencia.
[3] Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona.
[4] Department of Radiology. Hospital Universitario Doctor Peset. Valencia. Spain.

A portion of this paper represents the first author's doctoral thesis research.

Corresponding author: Estanislao Arana.
Department of Radiology. Hospital Casa de Salud. Manuel Candela, 41.
E-46021 Valencia. Spain. Fax: +34-963692635. E-mail:aranae@uv.es.

Abstract

The objective of this paper is to compare the performance of two predictive radiological models, logistic regression (LR) and neural network (NN), with five different resampling methods. One hundred and sixty-seven patients with proven calvarial lesions as the only known disease were enrolled. Clinical and CT data were used for LR and NN models. Both models were developed with cross validation, leave-one-out and three different bootstrap algorithms. The final results of each model were compared with error rate and the area under receiver operating characteristic curves (Az). The neural network obtained statistically higher Az than LR with cross validation. The remaining resampling validation methods did not reveal statistically significant differences between LR and NN rules. The neural network classifier performs better than the one based on logistic regression. This advantage is well detected by three-fold cross-validation, but remains unnoticed when leave-one-out or bootstrap algorithms are used.

KEY WORDS*:* Skull, neoplasms; Statistics, logistic regression; Neural networks; receiver operating characteristic curve; Statistics, resampling.

JEL Classification: C13, C14.

Introduction

Predictive models have been extensively studied as supporting diagnostic aids to radiology on a variety of diseases (1). Among these methods, neural networks are developed for radiological purposes. They are based on a parallel architecture with several layers. Each layer receives input only from the directly preceding one and adjacent layers are fully connected. The utility and flexibility of neural networks arises from the application of learning algorithms that allow the network to construct the correct weights and, hence, the desired function, for a given set of observations. Although several algorithms have been discussed in the literature, the most commonly employed is the backpropagation of errors algorithm.

Logistic regression is a nonlinear regression technique that has proven to be very robust in a number of medical domains and is acknowledged as the statistical analysis of choice for predicting dichotomous outcomes (2).

A standard procedure for evaluating the performance of a model would be to split the data into a training set, a cross-validation set (used to determine the stopping point to avoid over-fitting, and/or used to set additional parameters, such as weight-elimination), and a test set. The test set is a set of examples, not previously shown to the neural network, and only used to assess the performance (generalization) of a fully-specified classifier. Regression models should generally use a training and a test set as well (3). In medical settings, frequently due to the small amount of data available for training these models, new cases are rarely found to be tested. For these reasons, validation procedures with resampling techniques are usually employed. These techniques use part of the data set to train and to validate these models. Recent radiological papers dealing with neural networks have discussed these issues. However, many researchers are using such models without validating the necessary assumptions (4). Models developed this way are unlikely to stand the test validation on a separate patient sample (3). The main NN drawbacks are that usually they are not matched with statistical techniques of reference and they are used in small samples (2). The associated risk of over-fitting on noisy data is of major concern in neural network design (5), although logistic regression methods can suffer from the same problems as neural networks (2).

Calvarial lesions are often found during CT imaging of the brain with no specific symptoms. The signs for their characterization are based on imaging studies, specially CT. A diagnostic model could help the radiologist with these uncommon lesions (6). The purpose of this work was to study the ability of five resampling methods, leave-one-out and bootstrap, for validating logistic regression and neural network models to classify calvarial lesions and compare their results.

Material and methods

Population

A complete discussion of the population and methodology has been previously presented (6). As part of our review, 167 patients with calvarial focal bone lesions were reviewed for a four-year-period from 4,012 head CT scans. The lesions were analyzed in detail and reviewed by two of us by consensus (EA and LM-B). All patients were examined with at least two plain radiographic projections and CT. CT sections were obtained with the window width and level settings that best allowed evaluation of soft-tissue structures and bones. The setting on each scanner was individualized for every patient.

There were 74 men (44.6%) and 93 women (55.4%) with an age range of 0.5 to 81 years (31.1 ± 25.5 years, mean ± SD). The total number of benign lesions was 122 (73.1%), with an age range from five to 80 years (26.1 ± 22.9). There were 45 malignant ones (27.0%) with an age range from 0.5 to 84 years (55.7 ± 18.6).

Explanatory diagnostic variables

Nineteen morphological imaging characteristics as well as anatomic and demographic data were evaluated without knowledge of the final diagnosis (Table 1). All findings were recorded for all patients in a spreadsheet and used for both the LR model and the NN analysis. When a single feature had two or more findings in the same lesion, the most severe form was the one recorded. There was no missing data. Lesions were divided into benign (0) and malignant (1). Out of the 19 explanatory variables, three were continuous (1,4,8), one was quantitative discrete (5), two were ordinal (11,12) and the rest were qualitative. For these latter, we used a 1-out-of C code where a variable with C categories is converted in C Boolean inputs, each of which is high for a certain category; eventually 43 explanatory variables were used.

Logistic regression

Logistic regression equation assumes that the expected probability of a dichotomous outcome is

where the X_i are variables with numeric values (if dichotomous, they are, for example, 0 for false and 1 for true) and the b_i are the regression coefficients which quantify their contribution to the probability. Logistic regression has proven to be an effective way of estimating probabilities for dichotomous variables, in this case the probability of benign (0) or malignant (1).

We fit the logistic regression model by a Newton-Raphson type iterative algorithm implemented in MATLAB 4.0 (MathWorks, Inc.; Natick, MA). The nineteen explanatory variables listed in Table 1 were re-codified as 43 input variable and all of them used in the b_i parameters fitting. The estimation procedure also provides an measurement of the estimated coefficients standard errors. This allows defining confidence intervals (CI) for the coefficients and quantifying if they are significantly different from zero. The value of a standardized coefficient is an indication of the relevance of the corresponding explanatory variable. The usual output of a logistic regression estimation is the odd-ratios list (i.e., the values e^{b_i}) and the CI for them, derived from the coefficients CI (also by using the exponential function).

Neural network

The developed NN had three layers, with a feed-forward architecture, being trained by the back-propagation algorithm with the sigmoid activation function (7). The structure of the NN included 43 input units, which were the same as in the LR model. No well-established theoretical method exists for designing an ideal NN, and the optimal number of hidden nodes and iterations is unknown (7,8). The number of hidden neurons was chosen among those with a satisfactory adjustment of the NN. The number of tested hidden neurons ranged from 10 to 30 and a satisfactory trade-off between low miss-classification rate and complex design was obtained with 15 hidden neurons. Two stopping rules were implemented, one based on the number of iterations (the maximum allowed number was 500) and the other based on a lower bound for the mean squared error (MSE) that indicates how well the outputs are calibrated to their targets (this bound was equal to 0.03). In fact, the effective stopping rule was always the first one in all the training procedure run. The main drawback of these stopping criteria is the possibility to find a local optimum instead of the global optimum value. This is a common problem for a wide variety of multivariate optimization procedures. Some techniques including random perturbations of intermediate solutions (as simulated annealing; see for instance Ripley

(9)) are specially designed to lead to a global optimum. We limit ourselves to standard implemented training processes in the neural network MATLAB toolbox.

The main concern in a design with an excessive number of hidden nodes is over-fitting. The number of hidden nodes determines the complexity of the functions represented by a NN: as this number increases, the function is more complex. If a net contains too many hidden nodes, this net can learn the training set so perfectly that even a zero training error could be achieved (then we say that the net over-fits the training set), but this net will usually have a big generalization error in independent test sets. This situation will appear because, at least in theory, we take the *best* function among too many flexible classes of functions. In practice, only one function is taken when a NN is trained: the optimization procedure is limited by the specific training algorithm and stopping rules implemented (number of iterations, required level of precision, etc.). So, not only the number of hidden nodes determines the complexity of the NN, but also the training algorithm. In our case, we only use 500 iterations in the training procedure. Since not every function with 15 nodes can be learned in 500 iterations, the class of actual possibly learned functions is not so big and, as a consequence, over-fitting problems may not appear.

Comparison and performance

One of performance measures of a classification rule is the probability of miss-classifying a new observation, assuming that a case is assigned to a class if the classification rule gives it a probability higher than 0.5 to belong to that class. A naive estimator of this probability is the apparent classification error or error rate, defined as the number of incorrectly classified cases in both classes divided by the number of total cases. This estimator is optimistically biased because the same cases are used to fit the classifier and to compute the error rate.

Another measure of performance is the area (A_z) under the receiver operating characteristic curve (ROC). The continuum output given by LR and NN was compared with their correct values to obtain the ROC curves. ROC curves measure predictive utility by showing the trade-off between the true-positive rate (*sensitivity*, probability of correctly classification a positive case) and the false-positive rate (*1-specificity*, probability of incorrect classification for a negative case) inherent in selecting specific thresholds on which predictions might be based. The area under this curve represents the probability that, given a positive case and a negative one, the classifier rule output will be higher for the positive case and it is not dependent on the choice

of decision threshold. This way is less dependent on the frequency of malignancy in the population, and allows considering the sensitivity and specificity of the model at various probability levels. In this study, the area under the ROC curve was obtained by plotting sensitivity versus 1-specificity for each possible predictive score cut-point, and summing the areas of the created trapezoids. Statistical differences and confidence intervals between the NN and LR outputs were compared with a two-tailed, nonparametric approach according to the method described by Hanley and McNeil (10,11).

Validation

Pure naive validation methods use all the cases to build the models and also to validate them. As the estimated model has been fitted to the idiosyncrasies in the training sample, the validation based on the same sample tends to under-estimate the probability of misclassification. To estimate the performance of a classifier, a validation method with lower bias is preferred. For these reasons resampling methods are used to depict its future prediction accuracy, but also for choosing a classifier from a given set (model selection), or combining classifiers.

Cross-validation with random sub-sampling

In *k*-fold cross-validation, the data set is randomly split into *k* mutually exclusive subsets (the folds) $D_1,D_2,...,D_k$ of approximately equal size. The classifier-rule is trained and tested *k* times. The cross-validation estimation of accuracy is the overall number of correct classifications, divided by the number of instances in the data set. As we have a moderate size sample (n=167), we developed a cross validation with *k*=3. For each division of the sample, a model is developed with n_1=112 and tested in the rest n_2=55. This way, we are in the ideal situation of having an independent sample to test the model. This process is repeated 20 times, with randomly chosen training and testing sets giving up 60 unbiased estimates of discriminant ability. As the test samples are independent of the training data, the results derived from this 3-fold cross-validation are reliable. The drawback for obtaining this reliability is that a third of the data in the model estimation phase is lost. Error rates are the average of the 20 resampling process.

Leave-one-out

Several validation methods are available if one cannot afford loosing a significant part of the sample in the estimation step. One of them is the leave-one-out and others are based on bootstrap principles. In a sample of size *n*, leave-one-out is *n*-fold cross-validation. According

to this method, all the database (n=167) but one patient are used to develop the classifiying models. Then the LR or NN model is tested on the case that is left out. The same process is repeated so that every pattern of the data is left out once.

Bootstrap

According to the bootstrap method, training set (a bootstrap sample) is generated by sampling with replacement n times from the available n cases. A diagnostic model is trained on the bootstrap model and then tested on both the bootstrap and the original set and accuracy is measured twice. The difference between both rates of misclassifications reflects the optimism of the naive apparent error rate. The same process is repeated B times and the average of those differences is taken as a global measure of the optimism. The estimations developed also included two different bootstrap algorithms as described in (12):

- apparent error rate + optimism (bootstrap-1).
- error rate 0.632 bootstrap (bootstrap-2).
- error rate 0.632+ bootstrap. This combines the "leave-one-out bootstrap" with a measure of over-fitting (bootstrap-3).

The number of B sub-samples generated was 100 and results were obtained with the *bootpred* routine written in S-plus and described by Efron and Tibshirani (13).

Results

Logistic regression and neural network fit

Only some numerical results from the estimation phase are reported here because our main objectives were the comparison among different resampling procedures. Table 2 shows the most relevant variables in the logistic model fit, sorted by the standardized coefficient values. Odds ratios and 95% CI for them are shown. It should be noticed that only the first two variables (*Age* and *Mixed blastic permeative character appearance*) have CI for odds ratios without including the value one. The results obtained in the NN and LR fitting, which are listed below, allowed us to state that the designed NN (with 15 hidden nodes and 500 iterations) does not over-fit the data, because logistic regression (LR) presents a lower training error than NN fit, but LR has greater generalization error.

Pure naive models

The apparent error rates for LR model was 0.0240, with an Az 0.9993 ± 0.0036. The NN showed an error rate of 0.0599 with an Az 0.9505 ± 0.0285. These results reveal the over-fitting of the models without resampling methods.

Cross-validation with random sub-sampling

ROC Az with their confidence interval (95%) for the different resampling methods are showed in Figure 1 and Table 3. Both models, particularly NN, obtained the smallest variances among the other resampling methods (Fig.1). The logit model performed poorer than NN with higher error rates of 0.1916 versus 0.1377 ($p<0.0001$), with A_z 0.8103 versus 0.8854 ($p<0.001$) (Fig. 2). The p-values reveal larger differences when error rates are compared instead of A_z .

Leave one out

The LR using the leave-one-out method performed significantly better than the cross-validation in terms of error rate ($p<0.01$). Although, the NN obtained higher A_z and lower overall error rates than the logit model, there were no statistical differences. (Table 3). Areas under ROC curves presented markedly larger variances than cross validation, although smaller than seen with bootstrap (Fig.1).

Bootstrap

There were no statistical differences between the LR and NN models validated with the three bootstrap algorithms either in error rates or in A_z's. The third bootstrap approach (.632+ bootstrap) gives similar results to 3-fold cross-validation (Fig. 3). The opposite happens when results from bootstrap-1 are examined (it shows that RL performs marginally better than NN) and bootstrap-2 indicates advantages for NN in terms of the global measure A_z, but not in terms of error rate (Table 3). Variance of A_z estimations was progressively greater in every bootstrap algorithm showing profound differences with cross-validation algorithms (Fig. 1).

Discussion

There is a great interest in comparing neural networks and classifier rules in medical applications. Ideally, we would train the NN and LR models with a larger data set and apply the trained models to an entirely different data set to evaluate its performance. However, we were not able to split our database into completely separate training and testing data sets because the number of confirmed lesions is currently small, although, to our knowledge, it is the largest series in the literature. Future research is needed to compare trained models on completely

separate data sets. Although some theories have been presented on sampling methods, these are still in their infancy (14).

The objectives of these studies on predictive models should be clarified because the approaches are different if we want to find the "gold standard predictive" model, or a model whose prediction error is the lowest. The classification performance of stochastic models, such as LR and NN, however, depends on the estimation techniques. Thus, error rate is optimistically biased and A_z is not. This may be responsible of the greater differences between LR and NN when they are compared in terms of error rates.

In the medical practice, data set size is always finite and usually smaller than desired. The main drawback of *k*-fold cross-validation is that it makes inefficient use of the data: a third of the data set is not used to train the classifier rule. Previous studies on this method have shown that, as the training sample size increased, so did the NNs predictive accuracy (15). One may expect that these error rates would decrease and, respectively A_z's would increase, when the whole sample is used to estimated both RL and NN. To extrapolate these results from our sample size of 112 to 167, the relevance of the numerical values is more qualitative (NN clearly surpass RL) than quantitative. So, the error rates obtained may be far from the reliability with larger samples.

According our results, leave-one-out has been described as an almost unbiased method but with high variance (16). Leave-one-out gave similar variance to the other resampling methods but cross-validation. The leave-one-out procedure assures that, regardless of the sample size, the relevant observation would not be in the pseudo-optimal solutions more than once (14). This resampling technique can easily produce significantly different results in NN settings, depending on the training-stopping criterion (15). A theoretical study suggested that cross-validation and leave-one-out do not offer significant improvement over the apparent error, whereas the improvement given by bootstrap is substantial (17). Nevertheless, these comparisons were carried out using simulated data and the root mean squared error for performance measuring, instead of real data and ROC analysis. Cross-validation can be very sensitive to the specific sample splitting. Furthermore, if the specific test set is given and the data is sparse and noisy (as in medical settings), test of predictive reliability may not reflect a good picture of sample variability, or potential changes in model specification (18).

Neural networks validated with cross-validation in radiological diagnoses have shown protean features. Theory has been corroborated with real data as the training sample size increased, so did the network's predictive accuracy, e.g. in ventilation-perfusion imaging

(15,19). However, in other fields as focal bone lesions, the performance appeared to be more strongly related to how radiographically distinctive each pathologic type is rather than the number of cases available (20). So, simulation studies concluded that cross-validation gives a nearly unbiased estimate of error, but often with unacceptably high variability, particularly if the database is small (21). In contrast, our results showed that cross-validation obtained the smallest variance among the different algorithms. The reason for our small variance compared with Tourassi et al (15) and Efron (21) is that we randomly repeated 20 times the 3-fold cross-validation procedure obtaining 60 observations of cross-validation error to be averaged, further more than they did. Efron (21) suggested that different resampling methods applied to practical situations could give different answers. However, in more recent papers comparing resampling methods on radiological data, Tourassi (15) and current work, the results were very similar among them.

Bootstrap techniques in radiological diagnosis have only been described in this latter work (15). In 1997, Efron & Tibishirani (12) proposed the ".632+" estimator, which combines the "leave-one-out bootstrap" with a measure of over-fitting. In extensive simulations it has shown to be the best-performing bootstrap and offer some gains over cross-validation. Similar results appear in our study: ".632+" was the only bootstrap based procedure indicating similar results as those obtained by 3-fold cross validation. The available asymptotic results show its validity for a large number of linear, nonlinear and even nonparametric regression problems. In contrast to the Bayesian approach, no distributional assumptions (e.g., normal errors) have to be specified. For a large sample size and a small B value, bootstrap does not ensure that all the relevant observations will be deleted at least once. The advantage is that in small size samples and with a relatively large B value, bootstrap algorithms may capture variability of sample data better than the leave-one-out procedure (14). Particularly in NN settings, bootstrap yields rather reliable estimates of the variance even in small sample situations, models where the distribution of residual depends on the input and, in addition, it is more robust if the selected model is incorrect (13). Therefore, it does not need to assume normality or symmetry in the data.

A more practical question, which should be considered as well, is whether the bootstrap is worth the extra computer time required. It is important to keep in mind that each bootstrap iteration requires a run of the algorithm, and it seems unlikely that this can be improved upon. In our work, all bootstrap routines (with B=100) were the second longest after leave-one-out (n=167) with the NN (overall duration, respectively, 115'41'' and 180'10''). This is just a

manifestation of ‘Occam’s razor’ which states that complex models should not be preferred to simpler ones (22).

We agree with a previous paper dealing with the uncertainty of choosing resampling methods on neural network’s design (15). Although they argued that if the estimates of resampling methods is very similar, we can be confident about the classifier rule performance. However, other types of neural networks may show different performances due to the many parameters involved in neural network development, i.e., the leave-one-out and the training-stopping criteria. Frequent training problems exist and may be difficult to address even with state-of-the-art resampling methods. The required number of training cases clearly depends on the difficulty of the decision task, the number of input, hidden and output nodes (and also the number of weights).Thus, the information provided to the network is crucial. Reinus et al. (20) have proven improved performance of their NN’s design using a greater number of input units. This means that the imaging parameters were most explicitly defined than in other neural networks with less input features.

Nevertheless, one work which evaluated forecast on financial data found that the variation due to different resampling (i.e., splits between training, cross-validation, and test sets) is significantly larger than the variation due to different network conditions (such as architecture and initial weights) (18).

Summarizing the conclusions, neural network classification rule is preferred to logistic regression in the diagnosis of focal calvarial lesions. This advantage is well detected by three-fold cross-validation, but remains unnoticed when leave-one-out or bootstrap algorithms are used. However, both leave-one-out and “.632+” bootstrap slightly indicate the superiority of NN over LR.

Acknowledgment

We thank to John Boone, Ph.D., Professor of Radiology, University of California Davis Medical Center for his critical review and suggestions to this manuscript and the two reviewers for their ideas to improve our work.

References

1. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology* 1993;187:81-87.

2. Tu J. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49:1225-1331.

3. Harrell F, Lee K, Matchar D, Reichert T. Regression models for prognostic prediction: advantages, problems and suggested solutions. *Cancer Treat Rep* 1985;69:1071-1077.

4. Christy P, Tervonen O, Scheithauer B, Forbes G. Use of neural network and a multiple regression model to predict histologic grade of astrocytoma from MRI appearances. *Neuroradiology* 1995;37:89-93.

5. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Computation* 1992;4:1-58.

6. Arana E, Martí-Bonmatí L, Paredes R, Bautista D. Focal calvarial bone lesions. Comparison of logistic regression and neural network. *Invest Radiol* 1998;33:738-745.

7. Penny W, Frost D. Neural networks in clinical medicine. *Med Decis Making* 1996;16:386-398.

8. Miller A, Blott B. Review of neural network application in medical imaging and signal processing. *Med Biol Eng Comput* 1992;30:449-464.

9. Ripley BD. *Pattern recognition and Neural Networks*. Cambridge University Press, 1996.

10. Hanley J, McNeil B. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 1982;143:29-36.

11. Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839-843.

12. Efron B, Tibshirani R. Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. *J Am Stat Assoc* 1997; 92:548-560.

13. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. London: Chapman & Hall; 1993.

14. Ziari H, Leatham D, Ellinger P. Development of statistical discriminant mathematical programming model via resampling estimation techniques. *Amer J Agr Econ* 1997;79:1352-1362.

15. Tourassi GD, Floyd CE. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. *Med Decis Making* 1997;17:186-192.

16. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Technical report. Computer Science Department. Stanford University*. 1995.

17. Gong G. Cross-Validation, the Jacknife, and the Bootstrap: Excess error estimation in forward logistic regression. *J Am Stat Assoc* 1986;81:108-113.

18. LeBaron B, Weigend AS. Evaluating neural network predictors by bootstrapping. Technical report. Madison,Wisconsin. University of Wisconsin. 1995.

19. Fisher R, Scott J, Palmer E. Neural networks in ventilation-perfusion imaging. Part I. Effects of interpretative criteria and network architecture. *Radiology* 1996;198:699-706.

20. Reinus WR, Wilson AJ, Kalman B, Kwasny S. Diagnosis of focal bone lesions using neural networks. *Invest Radiol* 1994;29:606-611.

21. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* 1983;78:316-330.

22. MacKay DJC. Bayesian model comparison and backprop nets. In: Moody J, Hanson S, Lippman R, eds. *Advances in neural information processing systems 4*. Morgan Kaufmann, San Mateo; 1992: 839-846.

Variables		Values
1	Age	Years
2	Gender	Male, female
3	First symptom noticed	Tumour, pain, headache, incidental finding, others
4	Symptoms length	Months
5	Number of lesions	Number
6	Bone	Frontal, parietal, occipital, sutures or fontanels
7	Centricity	Outer table, diploe, inner table, intracranial, extracranial
8	Maximal diameter	In millimeters
9	Shape	Circular, ovoid (image in plane of greatest diameter)
10	Character-appearance	Lytic permeative, lytic moth-eaten, lytic geographic, blastic, Mixed blastic permeative, mixed blastic moth-eaten, mixed blastic geographi
11	Expansivity	No, mild, moderate, severe
12	Edge definition	Poor, moderate, well
13	Lobularity	Lobular, smooth
14	Marginal sclerosis	No, partial, rind (<2 mm), band (>2 mm), no applicable
15	Periosteal reaction	No, yes (any form)
16	Matrix	None, ground glass, calcified, ossified and sequestration
17	Cortical involvement	Diploe, internal, external, both corticals
18	Form of cortical involvement	None, thickened, thinned, broken
19	Swell/mass	None, intracranial, subgaleal, both

Table 1. Variables recorded and their description.

Variable	Odds ratio	CI 95%
Age	1,117	1,036-1,203
Character-appearance		
Mix. blastic permeative	0,004	0,000-0,903
Periosteal reaction	0,041	0,001-1,250
Symptoms length	0,949	0,894-1,006
First symptom noticed		
Tumour	34,917	0,495-2464,392
Character-appearance		
Mix. blastic moth-eaten	0,009	0,000-2,770
Centricity		
Outer table	16,555	0,544-503,762
⋮	⋮	⋮ ⋮

Table 2. Most relevant variables included for the logistic regression model, $p<0.05$.

CI: confidence interval.

Algorithm	Logistic regression		Neural network	
	Error rate (CI 95%)	ROC Az (CI 95%)	Error rate (CI 95%)	ROC Az (CI 95%)
Cross validation	0.1916 (0.1810,0.2080)	0.8103 (0.7883,0.8323)	0.1377 (0.1240,0.1476)	0.8854 (0.8674,0.9034)
Leave one out	0.1377 (0.0854,0.1899)	0.8711 (0.7854,0.9567)	0.1198 (0.0705,0.1690)	0.8736 (0.7887, 0.9584)
Bootstrap-1	0.0958 (0.0539,0.1446)	0.8819 (0.7994,0.9644)	0.1737, (0.1162,0.2311)	0.8508 (0.7600, 0.9415)
Bootstrap-2	0.1377 (0.0840,0.1880)	0.8303 (0.7350, 0.9256)	0.1676, (0.1128,0.2267)	0.8351 (0.7408, 0.9294)
Bootstrap-3	0.1736 (0.1169,0.2321)	0.7809 (0.6766, 0.8852)	0.1676, (0.1128, 0.2267)	0.8361 (0.7420, 0.9302)

Table 3. Results of the logistic regression and neural network models with the different resampling methods in terms of overall error rate and area under the ROC curve (Az). CI: Confidence interval.

Figures

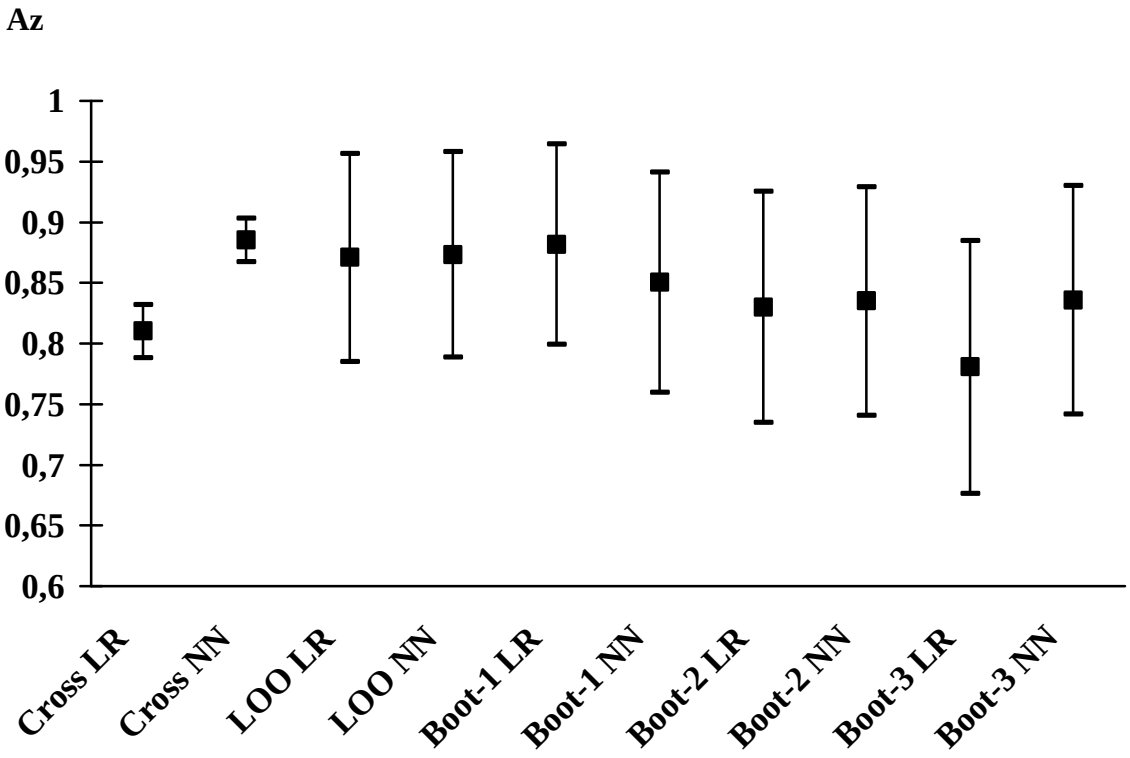

Figure 1. ROC Az with their confidence interval (95%) for the different resampling methods. Cross: cross-validation, LOO: leave-one-out, Boot-1: bootstrap-1, Boot-2: bootstrap-2, Boot-3: bootstrap-3.

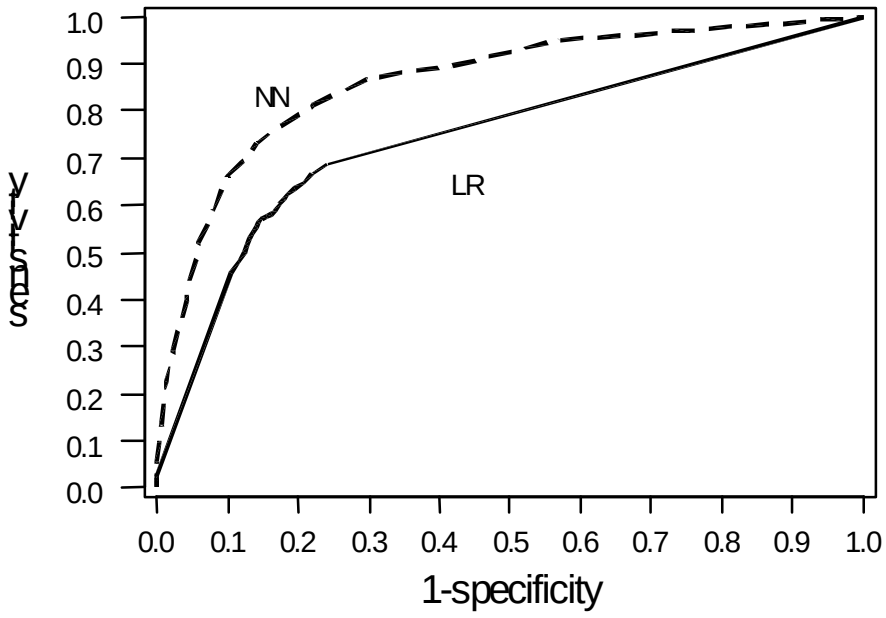

Figure 2: ROC curves for LR and NN based on 3-fold cross-validation. The reported curves are the average of the curves obtained in the 20 resampling processes.

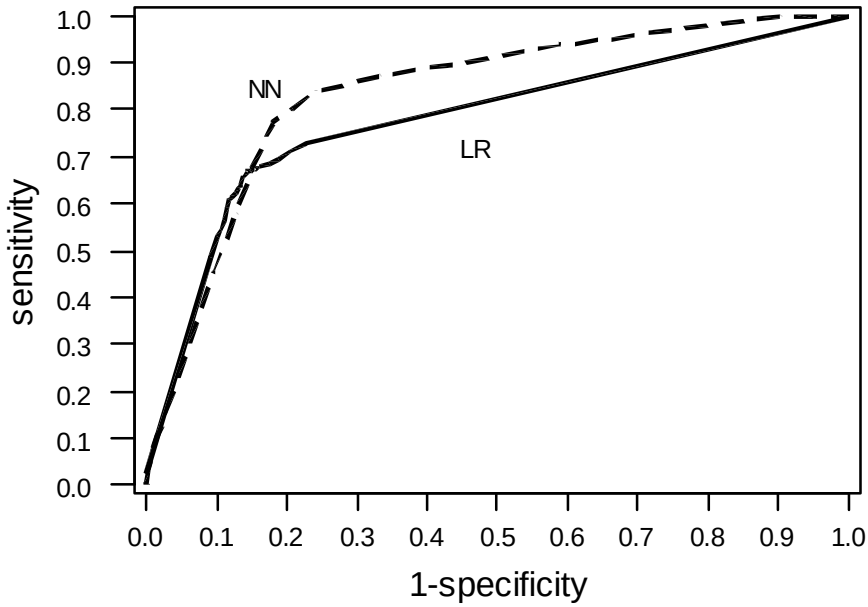

Figure 3: ROC curves for LR and NN based on .632+ bootstrap.