Models to predict cardiovascular risk: comparison of CART, Multilayer perceptron and logistic regression

Isabelle Colombet MD*, MPH, Alan Ruelland, MS*,
Gilles Chatellier MD*, François Gueyffier, MD, PhD**, Patrice Degoulet MD, PhD*,
Marie-Christine Jaulent PhD*,
*Medical Informatics Department, Broussais Hospital, Paris, France
** Clinical Pharmacology Department, Claude Bernard University, Lyon, France
colombet@hbroussais.fr

Abstract

The estimate of a multivariate risk is now required in guidelines for cardiovascular prevention. Limitations of existing statistical risk models lead to explore machine-learning methods. This study evaluates the implementation and performance of a decision tree (CART) and a multilayer perceptron (MLP) to predict cardiovascular risk from real data.

The study population was randomly splitted in a learning set (n=10,296) and a test set (n=5,148). CART and the MLP were implemented at their best performance on the learning set and applied on the test set and compared to a logistic model. Implementation, explicative and discriminative performance criteria are considered, based on ROC analysis.

Areas under ROC curves and their 95% confidence interval are 0.78 (0.75-0.81), 0.78 (0.75-0.80) and 0.76 (0.73-0.79) respectively for logistic regression, MLP and CART. Given their implementation and explicative characteristics, these methods can complement existing statistical models and contribute to the interpretation of risk.

Introduction

Current guidelines published for the management of the main cardiovascular risk factors (hypertension, hypercholesterolemia, type 2 diabetes) are based on a decision making strategy that uses a multivariate estimate of cardiovascular risk [1-3]. This strategy is supposed to lead to a more accurate identification of patients who will most benefit from the treatment of risk factors [4].

Any multivariate estimate of cardiovascular risk is currently based on the use of statistical models inferred from cohort data with methods such as logistic regression or Cox proportional analysis. Machine-learning methods are more and more explored and evaluated for risk prediction purposes in medical domains [5]. Few works have been published on the evaluation of machine learning methods to predict cardiovascular risk. Knuiman et al. find similar discriminative performance of a decision tree and a logistic regression model to predict coronary mortality in the Busselton cohort [6]. Lapuerta et al. uses a neural network to predict coronary risk from serum lipid profiles, taking into account censored data; the neural network showed a higher proportion of observations which are correctly classified compared with a Cox model [7]. The evaluation of prediction methods relies on the analysis of predictive performance of models. This performance is often incompletely assessed by such indicators as the proportion of correctly classified observations. Other aspects of methods implementation are not always addressed.

The objective of this work is to evaluate the implementation and performance of two machine learning methods (a multilayer perceptron and an inductive decision tree based on the CART algorithm) comparatively with a logistic regression model, in order to predict the risk of cardiovascular disease in a real database from the INDANA project (Individual Data Analysis of Antihypertensive Intervention Trials) [8]. This paper describes how these methods have been applied to the INDANA database and presents a comparison framework. The results are reported and discussed according to this framework.

Material and methods

The INDANA database

The INDANA database has been previously described [8]. Briefly, this database consists in the individual data of 10 randomized controlled trials designed to evaluate the preventive effects of antihypertensive drugs. In this study, we only used data from the control groups. Observations with missing data were mostly clustered by trial and were dropped from the original dataset. The final dataset consists in 15,444 subjects, described by several clinical characteristics and prospectively followed during at least 6 years for incidence of cardiovascular outcomes. Problems of heterogeneity of outcome and predictive variables measurements between trials are addressed by Gueyffier et al. [8]. The outcome considered in this paper is the 6-year incidence of the combined endpoint defined by occurence of myocardial infarction, stroke or cardiovascular death. It is represented in the dataset by a binary variable: occurrence (class 1) or no occurrence (class 0) of the outcome event.

Application of learning methods

Three learning methods were used to fit a prediction model to the data: logistic regression, a neural network and an inductive decision tree. This fitting process is hereafter described for each method.

Logistic regression

The reference model was built by forced entry of 10 variables followed by removal of the ones with no significant partial correlation (*R statistic*). The *SPSS v7.5.2F for Windows* (1997) statistical package was used for these analyses.

Neural network: NevProp (Nevada backPropagation)

We use a common feedforward backpropagation multilayer perceptron (MLP) simulator developped in the NevProp software package at the University of Nevada and freely available on the Internet [9]. The prediction method is based on the nonlinear weighted combination of input units (i.e. predictive variables) to predict one or more output units (i.e. outcome variable). The learning process is iterative and essentially consists in adjusting the weights to decrease the output error. The network was specified with one input layer (representing the ten predictive variables), one hidden layer (including ten hidden units) and one output layer (with one output unit representing a binary cardiovascular event). Several sensitivity analyses were performed to test how the prediction results could be influenced by the variations of learning parameters and to elicit the most optimized network. These parameters refer to the architecture of the network (number of hidden units), the method of internal validation (number of iterations and data-splitting processes), the options of data pre-treatment (i.e. normalization of inputs), the activation function for hidden units, and the "ScoreThreshold" used by the system to classify a case from its predicted probability.

Decision tree: CART (Classification And Regression Tree)

We use the software CART v3.6, developed by Salford Systems [10] and based on Breiman's original algorithm [11]. An inductive decision tree is essentially a set of rules represented by decisional nodes and leaves (i.e. terminal nodes) which are assigned to a class. The learning process consists in 1) selecting the most discriminative variable according to an impurity function to partition the data, 2) repeating this partition until the nodes are considered pure enough to be terminal and 3) pruning the resulting complete tree to avoid overfitting. Here again, sensitivity analyses were performed to elicit the best tree. Learning parameters refer here to the choice of the impurity function (i.e. Gini index), the internal validation method (split-sample, cross validation, bootstrap), the specification of prior probabilities and/or misclassification costs.

Sampling method

A split-sample strategy is used for application of the prediction methods. Randomly selected two thirds of the dataset are used to learn the prediction model (learning set: n = 10,296). The remaining third is used to validate the model (test set: n = 5,148).

All predictive models were optimized from a set of ten predictive variables (age, sex, systolic and diastolic blood pressure, serum total cholesterol, binary or multi-category smoking status, diabetes, left ventricular hypertrophy on EKG, body mass index).

Comparison framework

A comparison framework was defined to consider metrics other than just performance of models. Three types of indicators were assessed, based on intrinsic properties of the algorithms and on properties that are clinically useful for the defined task:

- *Implementation criteria* reflect the difficulty to optimally apply the method to new data. Three qualitative criteria are considered: 1) control of the learning time 2) representation of the predictive variables (are any transformation required?) 3) representation of the output result (is any decision threshold implicitly used, is it automatically defined by the system or is it user-defined ?)
- *Explicative performance criteria* reflect the extent to which the model explains by itself the prediction process. Three criteria are considered: 1) expressiveness of the outcome result (binary classification versus any other membership function) 2) report on the predictive variables implied in the decision and their relative importance 3) availability of a graphical representation to understand the model itself.
- *Discriminative performance criteria* reflect the ability of the model to separate high risk subjects from low risk subjects. Three criteria are considered: 1) the ROC curve and area under it (or c index) [12], 2) the sensitivity (i.e. true positive rate) and 3) the specificity (i.e. true negative rate, or 1 – false positive rate).

All the ROC curve analyses were performed using the RocKit software which takes as input a vector of predicted probability along with the observed event [13]. The logistic model and neural network were applied to the test set to obtain this input vector. In CART, these probabilities had to be extracted from the terminal nodes information given in the output. This extraction was done with an EXCEL macro.

Results

Reference logistic model

The reference logistic model takes into account seven out of the ten original variables. Table 1 describes these clinical characteristics in the database. Table 2 presents their predictive importance in the logistic model. The n-categorical variables are transformed for the model into n - 1 binary variables.

Table 1: Descriptive characteristics of the total population for diseased and non diseased people

Mean (SD) or %	With outcome (n = 891)	Without outcome (n = 14,553)
Age (y)	60.5 (9.7)	52.5 (9.5)
Sex (% males)	66%	52%
SBP* (mmHg)	174 (22)	161 (20)
DBP* (mmHg)	98 (10)	98 (8)
Diabetes (%)	2.1	1.4
Smokers-Sk (%)	36.3	29.7
ECG-LVH (%)	24.2	11.2
BMI* (kg/m^2)	26.9 (4.5)	27.3 (4.6)
TC (mmol/l)	6.4 (1.2)	6.3 (1.1)

*: SBP: systolic blood pressure, DBP: diastolic blood pressure, Sk: binary smoking satus; TC: total cholesterol; ECG-LVH: left ventricular hypertrophy at ECG; BMI: body mass index

Comparison of implementation criteria

Multilayer perceptron (MLP)

The learning time was not a major problem with the MLP (no more than few seconds) as far as we did not use any bootstrapping method for optimization of the learning process. The learning time is slightly influenced by several other learning parameters such as the number of iterations required, the number of cross validations, the pre-treatment of data by normalization.

Input data have to be numerical and this is the only requirement for the learning system to work. Several options of data standardization were explored and did not influence the performance of the model.

NevProp provides predicted probabilities to belong to class 1, and various global performance indicators like the total proportion of misclassified cases. No classification matrix is directly available to compute the sensitivity and specificity. During the learning process, a decision threshold (so-called "Score Threshold") is used to optimize the error at each iteration. The same threshold can also be applied to the final predicted probability to ultimately classify each case.

CART

The learning time with the *CART* software also depends on the validation options which are chosen to optimize the learning process: all re-sampling methods are time consuming (cross-validation, bootstrap and bagging).

CART can take as input either continuous or categorical variables. No distributional hypothesis is required for these variables. However, the tree structure relies on their binarization.

CART output provides a classification matrix, that allows to calculate the sensitivity and specificity. For a given case, the classification process checks which terminal node applies to the case. The output includes the following information on each terminal node:

- a probability which represents the membership degree of the terminal node to the class. This probability is computed, taking into account the relative frequency of each class in the terminal node and its relative size compared with the whole sample.
- the class assigned to the terminal node according to its probability and to pre-specified misclassification costs.

Table 2: Ranking of predictive variables: comparison of CART, MLP and logistic regression

Logistic Regression (coefficient/SD)	MLP (%)	CART (%)
Age (13,7)	Age (100)	Age (100)
Sex (9,6)	Sex (38.1)	SBP (59.6)
Sk-cat2* (5,4)	SBP (17.5)	ECG (26.5)
SBP (5,4)	ECG (5.7)	Sex (23.0)
TC (4,3)	TC (5.7)	Sk-cat (16.3)
ECG1*(4,0)	Sk-bin (3.4)	Sk-bin (15.2)
ECG2*(3,3)	DBP (1.4)	TC (14.5)
Sk-cat3* (2,4)	Diabetes (0.7)	DBP (13.9)
Diabetes (2,4)	Sk-cat (0.2)	Diabetes (1.3)
Sk-cat1* (2,0)	BMI (0.2)	BMI (0.9)
ECG4* (1,8)		
ECG3 *(1,2)		

*binary variable recoded from a multi-categorical variable

Comparison of methods' explicative performance

Each method analyzed reports an indicator reflecting the predictive importance of variables. NevProp uses an Automatic Relevance Determination (ARD) function to rank the importance of variables in predicting the outcome. In CART, an indicator of variable importance is computed according to information collected at each node. This information refers to the improvement of discrimination attributable to each potential test on the variables. Results on variable importance reported in the logistic model and in the models optimized with NevProp and CART were slightly different (Table 2). In CART, a graphical representation of the decision tree helps to understand the role of all predictive variables and interactions between them. No such graphical representation is available in NevProp.

Comparison of discriminative performance results

Taking into account the primary sensitivity analyses, specifications of models were as follows:

- NevProp: 10 input units, 10 hidden units, 30 splits for cross validation (on the learning set), 30 iterations for learning process, ScoreThreshold at 0.1.
- CART: Gini impurity function, split-sample validation, misclassification costs at 1 (the influence of misclassification costs on model performance is described in Figure 1)

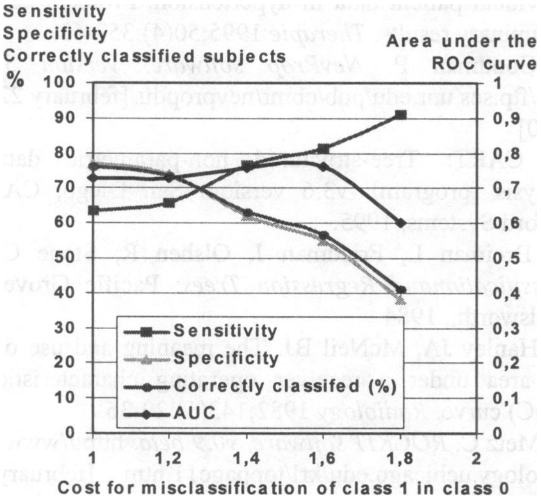

Figure 1: Variation of misclassification costs between 1 and 1.8 in CART: effects on model's performance

Table 4: Performance results for CART, MLP compared with the logistic model

	LR	MLP	CART
Correctly classified cases (%) in test set (n = 5148)	65.9%	76.0%	69.1
Area under ROC curve (95% CI)	0.78 (0.75-0.81)	0.78 (0.75-0.80)	0.76 (0.73-0.79)
AUCs difference with the logistic model (p value)		-0,9562 (p= 0.33)	-2,1864 (p = 0,02)

Table 4 describes the performance results of the models applied in the test set. ROC curves for CART and the MLP are not significantly different from the one obtained with the logistic model (Figure 2).

Discussion and Conclusion

This work comparatively evaluates the implementation and performance of two machine learning methods (a multilayer perceptron and an inductive decision tree based on the CART algorithm) by reference to a logistic regression model, in the real context of cardiovascular prevention. The predictive performance of CART is slightly lower than the performance of other methods However, we met several problems in the task of comparative evaluation.

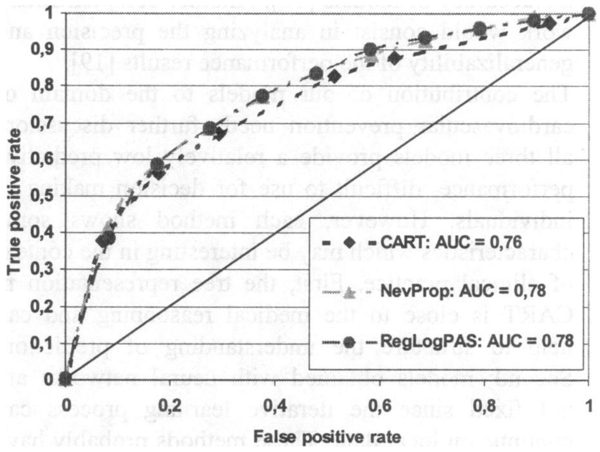

Figure 2: ROC curve for CART, MLP and logistic regression

First, at the implementation stage, we chose to evaluate the methods at their best performance, *i.e.* after optimization of the modeling specifications. This required to understand the meaning of each learning parameter and to test its influence on final results. Some qualitative standards should probably be clearly stated for implementation of these methods in order to make interpretable any kind of evaluation. An effort was already done in that direction in the NevProp manual [14]. A common environment is also still lacking for implementation and evaluation: Such an environment has been developed for UNIX and is not user-friendly [15].

Another difficulty was to define a common framework of indicators to evaluate the same type of results for each optimized model. This framework is based on an assessment of explicative performance and discriminative predictive performance by ROC analysis. Indeed, a risk prediction model can be considered as a diagnostic test and the ROC curve has been recommended as an appropriate measure of diagnostic accuracy by clinical epidemiologists [16]. While comparing methods, it is necessary to understand the semantic that underlies the output result of each method and to fit it into the common comparison framework. CART and MLP provide fundamentally different types of results. The extraction of predicted probabilities from CART output for the ROC analysis can be discussed. We chose an approach that has been already described and criticized [17]. Indeed, the probabilities available for each terminal node remain dependent of the tree's structure (namely its depth) and the interpretation of this probability may not be exactly the same as the one provided by the MLP or the logistic model. Moreover, we did not consider any measure of calibration which would provide along with discriminative performance a complete measure of the accuracy of models [18]. Another complementary work would consist in analyzing the precision and generalizability of the performance results [19].

The contribution of our models to the domain of cardiovascular prevention needs further discussion: all three models provide a relatively low predictive performance, difficult to use for decision making in individuals. However, each method shows some characteristics which may be interesting in the context of clinical practice. First, the tree representation in CART is close to the medical reasonning and can help to structure the understanding of prediction. Second; models obtained with neural networks are not fixed since the iterative learning process can continue on local data. These methods probably have the potential to complement existing statistical models and to contribute to the interpretation and presentation of risk in computerized decision support systems. Other machine-learning methods such as genetic algorithms, bayesian networks and support vector machines should also be explored.

Acknowledgements

We thank the collaborators of INDANA project (Individual Data Analysis of Antihypertensive Intervention Trials) for allowing us to access the INDANA database (Coope J, Cutler J, Ekbom T, Fagard R, Friedman L, Kerlikowske K, Mitchell Perry H, Pocock S, Prineas R, Schron E).

References

1. Prevention of coronary heart disease in clinical practice. Recommendations of the Second Joint Task Force of European and other Sbcieties on coronary prevention. *Eur Heart J* 1998;19(10):1434-503.

2. Jackson R, Barham P, Bills J, Birch T, McLennan L, MacMahon S, et al. Management of raised blood pressure in New Zealand: a discussion document. *BMJ* 1993;307(6896):107-10.

3. Unwin N, Thomson R, O'Byrne AM, Laker M, Armstrong H. Implications of applying widely accepted cholesterol screening and management guidelines to a british adult population: cross sectional study of cardiovascular disease and risk factors. *BMJ* 1998;317(7166):1125-30.

4. Grover SA, Paquet S, Levinton C, Coupal L, Zowall H. Estimating the benefits of modifying risk factors of cardiovascular disease: a comparison of primary vs secondary prevention. *Arch Intern Med* 1998;158(6):655-62.

5. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan B, Caruana R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997;9:107-38.

6. Knuiman MW, Vu HT, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *J Cardiovasc Risk* 1997;4(2):127-34.

7. Lapuerta P, Azen PS, LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Comp Biomed Res.* 1995;28:38-52.

8. Gueyffier F, Boutitie F, Boissel JP, Coope J, Cutler J, Ekbom T, et al. INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. *Therapie* 1995;50(4):353-62.

9. Goodman P. *NevProp software, version 3.* ftp://ftp.scs.unr.edu/pub/cbmr/nevpropdir [february 22 2000]

10. CART: Tree-structured non-parametric data analysis [program]. v3.6 version: San Diego, CA: Salford Systems, 1995.

11. Breiman L, Friedman J, Olshen R, Stone C. *Classificationand Regression Trees*: Pacific Grove: Wadsworth,, 1984.

12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.

13. Metz C. *ROCKIT software, v0.9 beta.* http://www-radiology.uchicago.edu/krl/toppage11.htm [february 22 2000]

14. Goodman P, Harrell Fj. *Neural networks: advantages and limitations for biostatistical modeling.* http://www.scs.unr.edu/nevprop [february 29 2000]

15. Rasmussen C, Neal R, Hinton G, van Camp D, Revow M, Ghahramani Z, et al. *The DELVE Manual, v1.1.* http://www.cs.utoronto.ca/~delve [february 29 2000]

16. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120(8):667-76.

17. Raubertas RF, Rodewald LE, Humiston SG, Szilagyi PG. ROC curves for classification trees. *Med Decis Making* 1994;14(2):169-74.

18. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-87.

19. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):515-24.