ELSEVIER

Journal of Clinical Epidemiology 57 (2004) 551–560

Journal of Clinical Epidemiology

Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism

Cornelis J. Biesheuvel[a,*], Ivar Siccama[b], Diederick E. Grobbee[a], Karel G.M. Moons[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center, P.O. Box 85500, GA Utrecht 3508, The Netherlands*
[b]*KiQ Ltd., De Lairessestraat 150, 1075 HL Amsterdam, The Netherlands*

Accepted 23 October 2003

Abstract

Objective: Genetic programming is a search method that can be used to solve complex associations between large numbers of variables. It has been used, for example, for myoelectrical signal recognition, but its value for medical prediction as in diagnostic and prognostic settings, has not been documented.

Study Design and Setting: We compared genetic programming and the commonly used logistic regression technique in the development of a prediction model using empirical data from a study on diagnosis of pulmonary embolism. Using part (67%) of the data, we developed and internally validated (using bootstrapping techniques) a diagnostic prediction model by genetic programming and by logistic regression, and compared both on their predictive ability in the remaining data (validation set).

Results: In the validation set, the area under the ROC curve of the genetic programming model was significantly larger (0.73; 95%CI: 0.64–0.82) than that of the logistic regression model (0.68; 0.59–0.77). The calibration of both models was similar, indicating a similar amount of overoptimism.

Conclusion: Although the interpretation of a genetic programming model is less intuitive and this is the first empirical study quantifying its value for medical prediction, genetic programming seems a promising technique to develop prediction rules for diagnostic and prognostic purposes.

Keywords: Logistic regression; Genetic programming; Prediction; Diagnostic research; Discrimination; Reliability

1. Introduction

In the past decade there has been an increased interest in medical prediction research to answer prognostic and diagnostic questions. Generally, such research aims to develop a so-called prediction rule to predict a particular outcome as accurate as possible, preferably with a minimum of information or predictors. In diagnostic prediction research the outcome includes the *presence* of a disease, and in prognostic prediction research the *future occurrence* of a certain event. With the increasing availability of electronic patient records the interest in medical prediction research will further increase because electronic records facilitate the application of prediction rules in medical practice.

The most widely used method to develop prediction rules or models in clinical epidemiology is multivariable logistic regression [1–7]. In the past decade, new methods such as classification and regression trees (CART) and neural networks have been introduced for this purpose. However, it has repeatedly been shown that both methods do not produce prediction rules that achieve higher predictive accuracy than rules developed by multivariable logistic regression [8–13]. Recently, the technique of genetic programming has emerged. Genetic programming is a search method inspired by the process of natural evolution, and may be used to solve complex associations between large numbers of variables [14–16]. This feature makes genetic programming also suitable for prediction research to estimate the mutual correlations between various predictors and the outcome.

Genetic programming is not restricted to any fixed model structure. Therefore, it may theoretically result in a model achieving higher predictive accuracy compared to a model obtained by ordinary logistic regression analyses. However, the flexibility a logistic model can also be increased by including cubic splines for continuous variables (rather than only the linear terms) and interaction terms, potentially enhancing the model's predictive accuracy [4,6,17]. However, this is not commonly done, as it often decreases the interpretability of such model.

* Corresponding author. Tel.: +31-30-2538633; fax: +31-30-2505480.
E-mail address: cbiesheu@umcutrecht.nl (C.J. Biesheuvel).

0895-4356/04/$ – see front matter
doi: 10.1016/j.jclinepi.2003.10.011

Like neural networks, genetic programming originates from the field of artificial intelligence and machine learning. But contrary to neural networks, genetic programming requires fewer prior restrictions to the structure of the model. Nevertheless, an often-cited disadvantage of both genetic programming and neural networks is the complexity of the developed prediction model ("black-box character"). Genetic programming has been used in medical research used for myoelectrical signal recognition, echocardiography, and medical imaging, but its value for medical prediction has not been documented yet.

Our aim was to compare genetic programming and multivariable logistic regression in the development of a diagnostic prediction model using empirical data from a study on diagnosis of pulmonary embolism (PE). We developed a prediction model using genetic programming and one using multivariable logistic regression, and compared both methods on their predictive ability in an independent data set. The feasibility to apply both prediction models in clinical practice is discussed, as well as the differences between genetic programming and neural networks.

2. Patients and methods

2.1. Patients: description of the empirical data set

For the present analysis, data were used from a prospective diagnostic study among 398 patients in secondary care of 18 years or older who were suspected of PE. As data are used for illustration purposes only, we refer to literature for details on the design and main results of the study [18–20]. Briefly, all patients underwent a systematic patient history and physical examination, followed by blood gas analysis, chest radiography, leg ultrasound, ventilation-perfusion lung scanning (VQ-scanning) and pulmonary angiography. Chest x-ray was considered abnormal, that is, indicative for the presence of PE, if it showed an elevated hemidiaphragm, a small pleural effusion, atelectasis, consolidation, or signs of heart failure. Leg ultrasound was considered abnormal if the femoral vein and/or popliteal vein were noncompressible. PE was considered present in the case of a high probability VQ-scan or abnormal angiogram after a nonconclusive VQ-scan, and absent in the case of a normal perfusion scan or normal angiogram after a nonconclusive VQ-scan. Of the 398 patients, 170 had PE (prevalence = 43%). All VQ-scans and angiograms were evaluated without knowledge of any other diagnostic information.

For our study, we *a priori* selected 10 candidate predictors, based on previous diagnostic studies [21–23]. These 10 predictors included eight patient history and physical examination predictors, that is, age, any coexisting malignancy, surgery within past 3 months, previous deep venous thrombosis, history of collapse, respiratory frequency, pleural rub, signs of deep venous thrombosis, and two from additional testing, that is, abnormal leg ultrasound and abnormal chest X-ray. We note that the data are used for illustration purposes and not so much to report the optimal model for prediction of presence of PE to be used in future practice.

2.2. Methods

The data set was split, randomly, in two parts: a derivation set of approximately 67% (165 patients) and a validation set of approximately 33% (133 patients). The derivation set was used for model development (both by the logistic regression and genetic programming method) and the validation set to test the validity of the two models. The aim of both methods was to develop a prediction model to estimate the presence or absence of PE as good as possible with a minimum of diagnostic tests (predictors).

2.2.1. Multivariable logistic regression

In the derivation set, we first fitted the overall model including all 10 predictors. In this modeling we did not simply include each predictor as linear terms. To enhance the flexibility of the logistic model and to obtain a more fair comparison with the (unrestricted) genetic programming, continuous variables (i.e., age and respiratory frequency) were included using cubic spline functions, both with four knots [4,6]. A reduced prediction model was obtained by selecting predictors with P-values $<.10$ using the Likelihood Ratio test. To further enhance a fair comparison with genetic programming, we quantified whether interaction terms between the selected predictors increased the model's predictive accuracy. As our aim was not so much to develop an easy applicable model for future practice, we analyzed many possible interaction terms. These, however, were not included all together but rather consecutively following the chronologic order in which predictors are measured in practice and following the order of contribution to the prediction [17]. Accordingly, we analyzed the interaction terms between chest x-ray with each of the selected history and physical examination predictors and of leg ultrasound with each of these predictors. In this, interaction terms of the selected history and physical predictors with the highest odds ratios were included first. The final model included all predictors and interaction terms with a P-value $<.10$.

Internal validation of the final model was performed using bootstrapping techniques [4,24]. Random samples were drawn with replacement from the derivation set with 100 replications, and the backward exclusion of the predictors including interaction terms was repeated within each bootstrap sample. Bootstrapping yielded an estimate of the overoptimism of the final model in predictive performance as expressed by the area under the Receiver Operating Characteristic (ROC) curve [4,24,25]. Furthermore, a shrinkage factor was derived from the bootstrap samples to recalibrate the model. To adjust the model for overoptimism (i.e., recalibration), this factor was used as a shrinkage factor or multiplier of the regression coefficients of the predictors in the final model [4,24–26]. The recalibrated model was applied to the validation set to estimate its discrimination and

reliability in an independent sample. All analyses were performed using S-plus 2000 (Insightful Corp., Seattle, WA).

2.2.2. Genetic programming

Genetic programming is a search method inspired by the biologic model of evolution [16,27]. It is an extension of the genetic algorithm first described by Holland [14] and Goldberg [15]. For the present analyses, we used the method of the OMEGA predictive modeling engine (KiQ Ltd., Cambridge, UK) [28] to search for a model that achieves optimal accuracy in predicting the presence or absence of PE.

In the genetic programming method by OMEGA, a prediction model is a mathematical formula, without inherent restrictions of complexity such as in logistic regression modeling, that uses all predictors (or a subset of these) as inputs. The building blocks of the formula are mathematical operators, chosen from a library of 20 operators. Each operator has two inputs and one output (Fig. 1, upper part). The output of the formula is a score, which is used to predict the presence of the outcome under study, a higher score indicating a higher probability. The fit of the formula also is expressed by the ROC area.

In the present study, first a set of 40 different prediction models was randomly created. This set consisted of 40 different mathematical formulas using different predictors. Then in an iterative process:

1. The fit of each model was determined by comparing the scores of the model with the observed PE frequencies in the derivation set
2. Various models were selected where models with a larger fit had a higher probability of being selected
3. Crossover and mutation between the selected models was performed, creating new models
4. These newly created models were moved to the next set of models and upon completing this set the next iteration started

As the mathematical formulas or models all consist of binary operators, they can be represented as a binary tree (see Fig. 1, upper part). To limit the amount of overoptimism, the trees were restricted to be no more than four levels deep, corresponding to a maximum of eight predictors.

In step 3, new models are created by simulating the natural processes of sexual recombination (here called crossover) between two chromosomes and mutation of DNA on a particular chromosome. In the context of genetic programming a prediction model or binary tree, as selected in step 2, can be compared with a chromosome. Here, crossover and mutation operate on the branches (i.e., parts of the formula having one or more predictors as input; see Fig. 1) and nodes of these trees. Given two selected models, crossover is realized by swapping branches between the two trees. The swapped branches are randomly chosen. A mutation occurs by exchanging a node or a branch in a tree with a randomly created substitute, also here the node or branch that is mutated is randomly chosen. This random process, in addition to the probabilistic nature of the selection step (step 2), prevents the search method from converging at a local optimum.

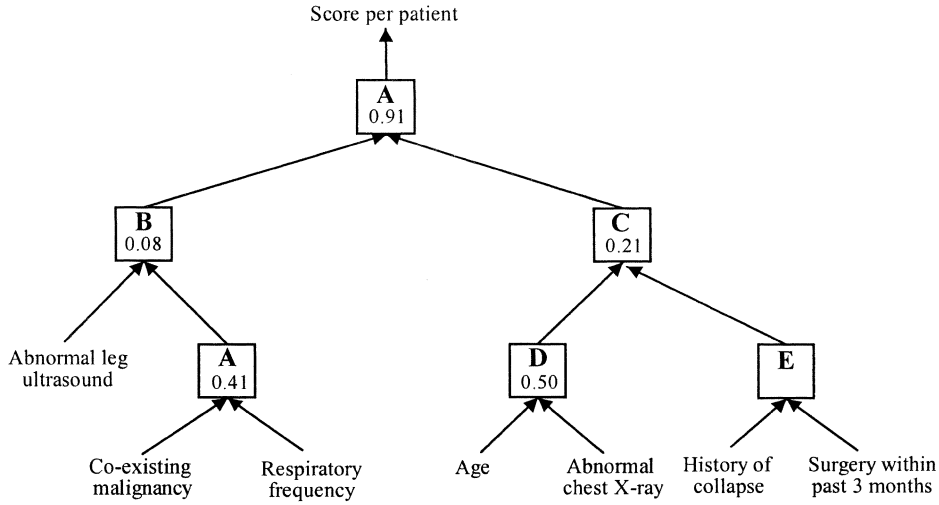
The iterative process as mentioned above was terminated when no significant model improvement was observed. The model in the population with the largest ROC area was then selected as the final genetic programming model. A model provides a score for each patient in the data set. This score was transformed to a probability of PE presence by linking them to observed or actual proportions of PE.

Similar to the logistic regression method, the entire process of model development for genetic programming, including predictor selection, was bootstrapped to estimate the amount of overoptimism in predictive performance (ROC area). Because the formula developed by genetic programming has a different structure and is more complex, there are no regression coefficients estimated compared to logistic regression models. Therefore, no shrinkage factor could be estimated, and the final model could not be recalibrated. The final genetic programming model was then also applied to the validation set to assess its discrimination and reliability in an independent sample.

2.2.3. Comparison of both methods

The two prediction models, obtained from multivariable logistic regression and genetic programming, were compared at their discrimination and reliability (calibration) in the validation set. Discrimination of both models was expressed by the area under the ROC curve. Reliability was evaluated by a graphical plot of the predicted probabilities vs. observed or actual proportions of the outcome and tested using the Hosmer and Lemeshow method [2]. For logistic regression, per patient the predicted probability was calculated using the recalibrated model, rank ordered, and divided into deciles. For each decile the mean predicted probability and observed proportion of PE was calculated. For testing reliability of the genetic programming model, the output scores were rank ordered and also divided into deciles. For each decile the average observed PE proportion was calculated. As genetic programming yields a score and not directly a predicted probability, we used the observed proportion of PE in the derivation set as predicted probability for the calibration curve. Hence, the calibration curve of the genetic programming model compares the observed proportion of PE in the derivation set (x-axis) to those found in the validation set (y-axis). To enable comparison of the unshrunk calibration curves of both models, for logistic regression the predicted probability of PE also was plotted against the observed proportion before bootstrap shrinking.

As said, bootstrapping of the developed genetic programming model did not yield a shrinkage factor such that the model could not be adjusted for overfitting (recalibration), in contrast to the logistic regression model. Hence, we also estimated the calibration curve of the original logistic model



Score category	Probability of pulmonary embolism
0.00-5.80	0.15
5.81-6.90	0.21
6.91-7.00	0.43
7.01-8.00	0.59
8.01-10.0	0.79

Fig. 1. The final model created by genetic programming, presented as a binary tree and output scores from the tree related to the observed proportion of the outcome. The nodes A–E represent the following binary operators, in which the parameters x (left arrow) and y (right arrow) are the inputs of each box:

$$A = 1 - p\sqrt{(1-x)} - (1-p)\sqrt{(1-y)}.$$

$$B = p\,f(x) + (1-p)\,f(y), \text{ where } f(x) = 2x - k\,(2x-1)^3 \text{ and } k = 0.593.$$

$$C = px + (1-p)\,y.$$

$$D = px^2 + (1-p)\,y^2.$$

$$E = \frac{1}{2} + \frac{1}{2}\sin(x^2 + \frac{1}{2}\pi y^2 - 1).$$

(before adjustment of overoptimism) and the genetic programming model for a fair comparison.

3. Results

3.1. Descriptives

There were no major differences in patient characteristics between the derivation and validation set (Table 1). PE was diagnosed in 42.6% of the patients in the derivation set, which was 42.9% in the validation set. Table 2 shows the univariable associations and distribution of the 10 predictors across patients with and without PE in the derivation set. "History of collapse" and "previous deep venous thrombosis" were the strongest predictors.

3.2. Model derivation

3.2.1. Logistic regression

The overall logistic model yielded a ROC area of 0.77 (95% CI: 0.71–0.83) and the reduced model, including eight predictors (Table 3), 0.76 (95% CI: 0.70–0.82). A restricted cubic spline transformation on age and respiratory frequency showed that the nonlinear terms for both predictors were far from significant (P-value $>.40$). Hence, age and respiratory

Table 1
Comparison of the outcome and predictors between the derivation and validation set

	Derivation set (n = 265)	Validation set (n = 133)
Pulmonary embolism present (%)	42.6	42.9
Age (years)	56.7 (17.8)[a]	53.8 (16.6)[a]
Any coexisting malignancy (%)	23.8	22.6
Surgery within past 3 months (%)	21.1	21.1
Previous DVT (%)	6.4	9.8
History of collapse (%)	7.2	9.0
Respiratory frequency (breaths/min)	19.7 (6.7)[a]	18.2 (6.1)[a]
Pleural rub (%)	14.7	18.1
Signs of DVT (%)	8.7	10.5
Abnormal chest X-ray (%)	40.4	39.1
Abnormal leg ultrasound (%)	24.2	21.1

DVT = deep venous thrombosis; min = minute.
[a] Mean (standard deviation).

frequency were analyzed as linear terms. Subsequently, we consecutively added the interaction terms between leg ultrasound with each of the selected six history and physical predictors. In this we first added the interaction term with collapse (as these showed the highest independent contribution to the prediction in predictive accuracy; see Table 3), followed by signs of DVT, pleural rub, and so on. The same was done for chest x-ray. All interaction terms were far from significant (P-value >.4) except for age with chest x-ray, which was borderline significant. However, this interaction term did not increase the discriminative power of the reduced model at all. Therefore, the model presented in Table 3 was considered as the final logistic model.

Bootstrapping estimated the overoptimism at 0.06 in ROC area. Hence, the internally validated ROC area of the final model became 0.70. The bootstrap shrinkage factor for the regression coefficients was 0.76. Table 3 (column 2) shows the adjusted association (shrunk coefficients) of each predictor retained in the final model with the outcome.

Table 2
Univariable association of each predictor with the presence of pulmonary embolism in the derivation set (n = 265)

Predictor	PE present (n = 113)	PE absent (n = 152)	Odds ratio (95% CI)
Age (years)	60.6 (16.5)[a]	53.7 (18.3)[a]	1.02 (1.01–1.04)
Any coexisting malignancy (%)	31.0	18.4	2.0 (1.1–3.5)
Surgery within past 3 months (%)	29.2	15.1	2.3 (1.3–4.2)
Previous DVT (%)	10.6	3.3	3.5 (1.2–10.2)
History of collapse (%)	14.2	2.0	8.2 (2.3–28.9)
Respiratory frequency (breaths/min)	21.2 (7.5)[a]	18.5 (5.9)[a]	1.06 (1.02–1.11)
Pleural rub (%)	19.5	11.2	1.9 (0.97–3.8)
Signs of DVT (%)	13.3	5.3	2.8 (1.1–6.7)
Abnormal chest X-ray (%)	49.6	33.6	1.9 (1.2–3.2)
Abnormal leg ultrasound (%)	33.6	17.1	2.5 (1.4–4.4)

DVT = deep venous thrombosis; min = minute.
[a] Mean (standard deviation).

Table 3
Association of each predictor in the calibrated final logistic regression model with pulmonary embolism

Predictor	Odds ratio	Regression coefficient	P-value
Age (per year)	1.01	0.011	.071
Surgery within past 3 months	1.57	0.45	.090
History of collapse	5.01	1.61	.002
Respiratory frequency (per breath/min)	1.05	0.044	.013
Pleural rub	1.81	0.60	.053
Signs of DVT	2.05	0.72	.061
Abnormal leg ultrasound	1.98	0.68	.006
Abnormal chest X-ray	1.46	0.38	.096
Intercept		−2.46	<.001

DVT = deep venous thrombosis.
Probability of pulmonary embolism in an individual patient: 1/(1 + exp − (−2.46 + 0.011*age + 0.45*surgery within past 3 months + 1.61*history of collapse + 0.044* respiratory frequency + 0.60*pleural rub + 0.72* signs of DVT + 0.68*abnormal leg ultrasound + 0.38*abnormal chest X-ray)).

3.2.2. Genetic programming

The final model developed by genetic programming included seven predictors (see Fig. 1). The ROC area of this model was 0.79 (95%CI: 0.73–0.85). Bootstrapping showed an estimated overoptimism of 0.07, decreasing the ROC area to 0.72.

3.3. Model validation

Applying the shrunk (adjusted) logistic regression model to the validation set yielded a ROC area of 0.68 (95% CI: 0.59–0.77), which was in good agreement to the ROC area estimated after bootstrapping. Application of the final genetic programming model to the validation set resulted in a ROC area of 0.73 (95% CI: 0.64–0.82), which was expected based on the ROC area after internal validation and significantly higher than the logistic model. Before shrinkage, both models showed similar calibration curves and predicted rather accurately over the entire range of observed proportions of PE (Fig. 2A and B). The Hosmer-Lemeshow test statistic was far from significant for both models (P-value >.50), indicating good reliability. The reliability of the recalibrated logistic model was obviously even better (Fig. 2C).

3.4. Model presentation

3.4.1. Logistic regression

The logistic model can be used in practice to estimate the probability of PE presence for individual patients in two different ways. First, one can use the formula as given in Table 3. With this formula one multiplies the patients' test results and corresponding coefficients, summing them and

antilog the sum. This method, however, requires a calculator. An easier method is using a nomogram as presented in Fig. 3. As an example of using this nomogram, a patient of 52 years of age (which corresponds to four points as determined from the "Points" scale on top of the figure), with recent surgery (4 points), history of collapse (15 points), breathing frequency of 20 breaths per minute (4 points), no pleural rub (0 points), no signs of DVT (0 points), normal leg ultrasound (0 points), and an abnormal chest x-ray (4 points) receives a "Total Points" score of 31. Using the lower two scales of the nomogram, this score corresponds to a probability of PE of approximately 0.8. The length of the line of each predictor in the nomogram also indicates the relative contribution of the predictor to the probability of PE.

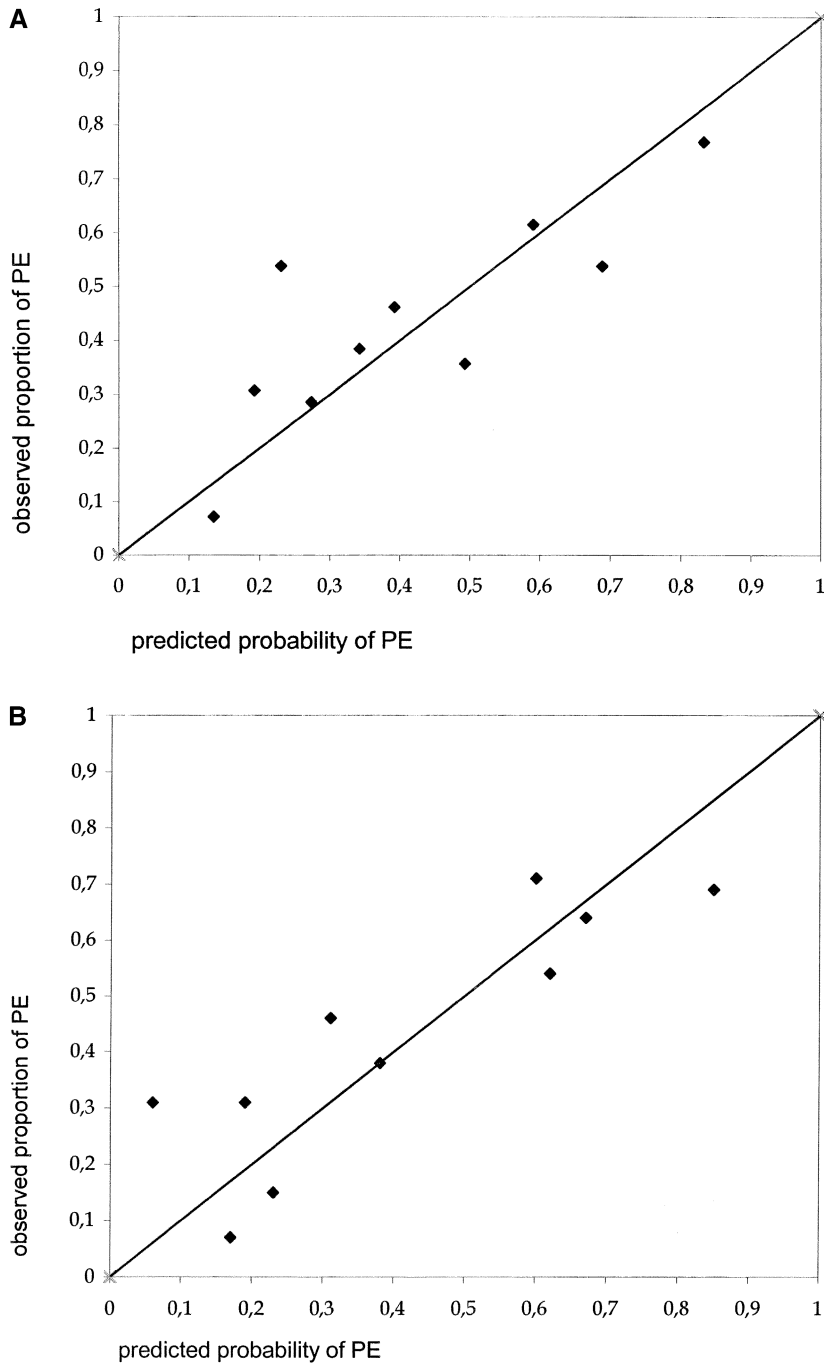


Fig. 2. (A) Calibration curve of the original (without application of bootstrap shrinkage) logistic regression model in the validation set. (B) Calibration curve of the genetic programming model in the validation set. (C) Calibration curve of the recalibrated (i.e., after application of bootstrap shrinkage) logistic regression model in the validation set.

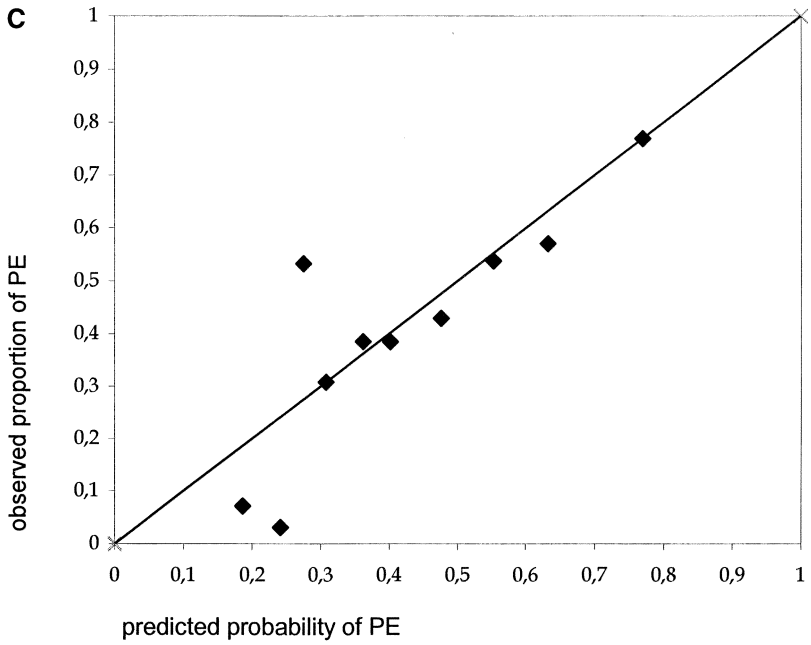


Fig. 2. Continued

3.4.2. Genetic programming

The model developed by genetic programming is shown in Fig. 1. The same patient as described above is taken as an example to calculate the probability of PE presence using the genetic programming model. The values of the predictors except for leg ultrasound (age = 52 years, surgery in past 3 months = 1, coexisting malignancy = 1, history of collapse = 1, respiratory rate = 20 breaths per minute, pleural rub = 0, abnormal chest X-ray = 1) are the inputs (x and y) for the boxes A, D, and E. The outputs of these boxes (using the formulas per box shown in the figure) plus abnormal leg ultrasound = 0, are then used as inputs of box B and C, which in turn, yield the input for box A. Box A results in the final score for the patient, which was 9.05. The score table in Fig. 1 shows that this score corresponds to a probability of PE of 0.79, which was similar to the probability obtained by the logistic regression model.

4. Discussion

To our knowledge, this is the first study to address the value of genetic programming for medical prediction purposes compared to the well-known and widely applied logistic regression technique. Given that the amount of overoptimism in discriminative value was similar for both models as estimated from the bootstrap, the discriminative value of the genetic programming model in the validation set was significantly larger than that of the logistic regression model. Before any form of recalibration or adjustment for overoptimism of the logistic model, the reliability in the validation set was similar for both models. However, the logistic model showed improved reliability after it was beforehand recalibrated through shrinkage. These results indicate that genetic programming offers a promising technique for prognostic and diagnostic prediction research, in particular, when the aim is to achieve optimal discrimination. To appreciate the results a few issues need to be addressed.

Because of the more complex structure of a genetic programming model, it does not provide regression coefficients or odds ratios that indicate the relative predictive contribution of each predictor. Therefore, logistic regression techniques remain the first choice when the primary goal of an epidemiologic study is to examine the (relative) strength of the association between risk factors and the outcome, such as in etiologic research. However, when the aim is to obtain an optimal prediction for individual patients in clinical practice, as is common for diagnostic and prognostic studies, researchers might search for more complex models as developed by genetic programming.

Due to the more complex nature of genetic programming, the final model is often less intuitive and interpretable ("black box" character). Some degree of insight can be given into the functioning of a genetic programming model by studying the sensitivity of the model to the used predictors, thereby identifying predictors that have a large influence on the output scores or discriminative value of the model. Also, insight can be gained by describing profiles of predictor values in the different score intervals and thus to specific risk categories. Nevertheless, the complexity of prediction rules developed by genetic programming may influence their use in medical practice. Application of a genetic programming model for prediction purposes in practice requires a personal computer equiped with the necessary software. Prediction rules developed by logistic regression can be easier

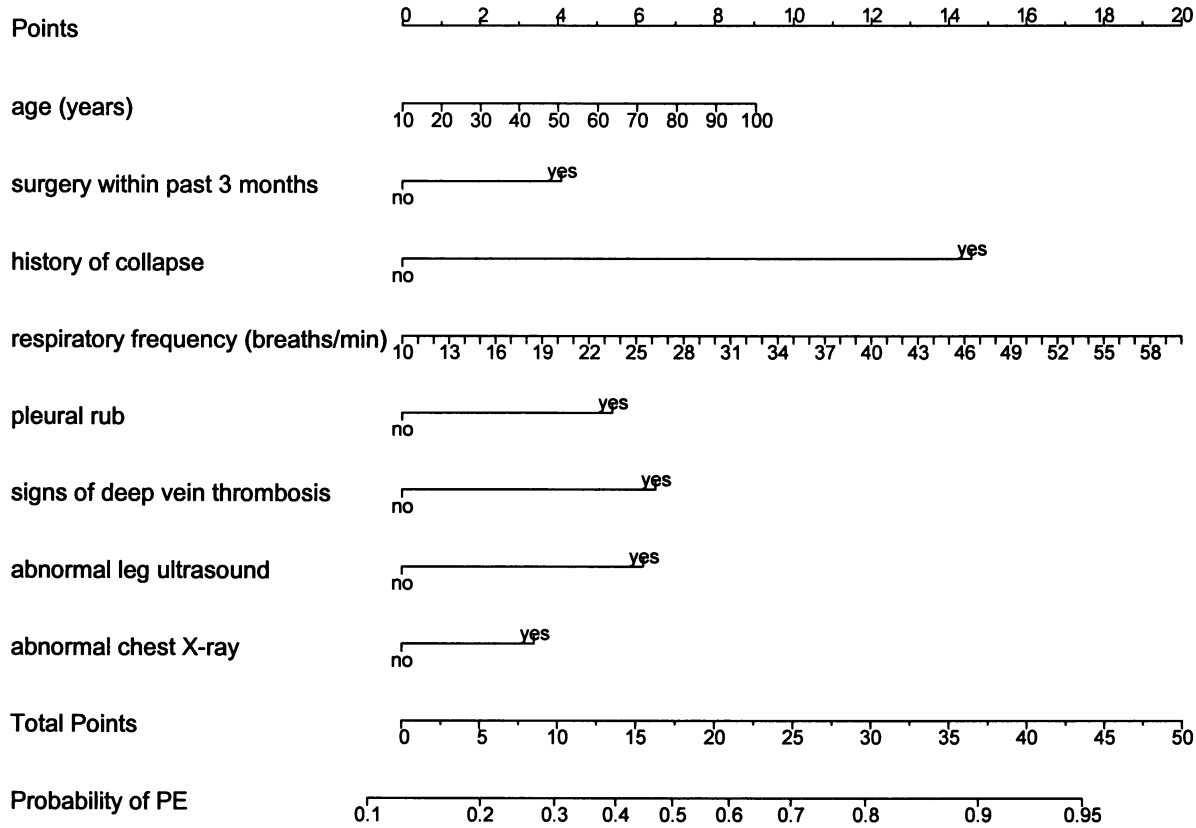


Fig. 3. Nomogram relating the predictors of the final model (Table 3) to the probability of pulmonary embolism. See text for instructions on how to use a nomogram.

disseminated because they directly yield a predicted probability by either using the regression coefficients and a pocket calculator or using a nomogram as presented in Fig. 3. However, with proceeding computerization in clinical practice and the rise of electronic patient records, this difference in applicability between both methods may become irrelevant.

The main advantage of the complexity and fewer restrictions of genetic programming is the possibility to create more flexible prediction models with better discrimination as was also exemplified in our analyses. This may be of even more importance in larger data sets in which complex interactions between predictors and outcomes may be present, although such interactions can also be modeled in a logistic model.

As a prediction model developed by genetic programming does not contain regression coefficients, a method of recalibration through shrinkage of these coefficients using data from the derivation set only (so-called internal validation) is not possible, in contrast to logistic regression. Furthermore, the patient score estimated by a genetic programming model is not linearly related to a probability of the outcome. To relate a score to probabilities, the observed frequency of the outcome must be estimated for different score intervals. These frequencies (probabilities) can be obtained from the data set from which the model is derived (derivation set). Subsequently, to determine its reliability, the genetic programming model should be applied to a validation set to compare the observed outcome frequencies per score interval from the derivation set to those observed in the validation set. To prevent overoptimism of the genetic programming model in practice, the probabilities from the validation set should then be used and presented as the probabilities that might be expected in future subjects. This method of recalibration for genetic programming models (as also done in our study) is appealing, but requires the use of a data set that is large enough to perform a split sample method.

An advantage of logistic regression is that a developed prediction model can easily be recalibrated using data from the derivation set only, for example, using shrinkage with the bootstrapping method (as done in our study), before it is to be applied to future patients. This is of particular interest when the available data to develop a prediction model is scarce. In our example study, *a priori* recalibration of the logistic model using internal validation techniques indeed improved the reliability of the model in "new" patients (see Fig. 2C). A similar technique of recalibration with the bootstrapping method using data of the derivation set only can also be

used for genetic programming, although without shrinking of coefficients. Instead, the probabilities per score interval are recalibrated by taking the average of the observed frequencies per interval obtained for each bootstrap model. This is a relatively time-consuming method because each bootstrap set requires the development of a new model. However, this method was performed and checked in our example study, and produced similar results to the validation method of recalibration. We only presented the latter results for reasons of clarity, and because frequencies (probabilities) from another patient (validation) set better reflect future practice.

Finally, genetic programming requires a number of parameters to be chosen by the researcher, such as the number of prediction models that are evolved in parallel (we used 40 models), the selection method, and the probabilities of crossover and mutation. Although the setting of these parameters requires some experience and certainly influences the speed of the search method, the final result was not very sensitive to these parameters; using other parameter settings did not result in other discrimination and reliability of the final model. The maximum depth for each tree is another parameter to be chosen by the researcher. It is chosen to limit the number of degrees of freedom, and is similar to the number of predictors that would be included in the final logistic regression model. Allowing for more degrees of freedom commonly results in a more overoptimistic prediction model.

Because neural networks are a well-known method and have also been used in the medical field to produce nonlinear prediction models, it is interesting to briefly discuss the differences between neural networks and genetic programming. Although the representation of the model created by genetic programming as shown in Fig. 1 may at first sight show similarities to a neural network model, it must be noted that only the representation in the form of a tree of the finally obtained (mathematical) formula is similar. In a genetic programming model, there are no hidden nodes. Instead, the boxes in the model (see Fig. 1) represent parts of the mathematical formula that are called operators. Because, in this case, the chosen operators are all binary, the representation is a binary tree. Indeed, a neural network as well is a presentation of a mathematical formula, but of a very specific and predetermined form. Given the chosen set of operators, the possible predictors to choose from, and restrictions on the number of degrees of freedom, the method of genetic programming is free to construct a more optimal mathematical prediction formula. The training period of a neural network (or the log-likelihood fit of a logistic regression), is, in the case of genetic programming, replaced by an iterative search process in which large numbers of possible models (mathematical formulas of different form, using different predictors) are evaluated in parallel, and which finally evolves into the most optimal solution.

In conclusion, using empirical data we demonstrated that a prediction model developed by the novel technique of genetic programming may have an increased discriminative power with comparable reliability, compared to a model developed by logistic regression. Although this is the first empirical study quantifying the value of genetic programming for medical prediction and more empirical studies are needed, it seems a promising technique to develop prediction rules for diagnostic and prognostic purposes.

Acknowledgments

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research\ (ZON-MW904-66-112).

References

[1] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.
[2] Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc.; 1989.
[3] Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–85.
[4] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
[5] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488–94.
[6] Harrell FE. Regression modeling strategies. New York: Springer-Verlag; 2001.
[7] Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56:337–8.
[8] Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. J Investig Med 1995;43:468–76.
[9] Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Medinfo 1998;9(Pt 1):493–7.
[10] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49:1225–31.
[11] Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17:2501–8.
[12] Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Fiedler RC, Granger CV. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. J Clin Epidemiol 2001;54:1159–65.
[13] Resnic FS, Ohno-Machado L, Selwyn A, Simon DI, Popma JJ. Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention. Am J Cardiol 2001;88:5–9.
[14] Holland JH. Adaptation in natural and artificial systems. Ann Arbor, MI: University of Michigan Press; 1975.
[15] Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison Wesley Publishing Company; 1989.
[16] Koza JR. Genetic programming III. Cambridge, MA: MIT Press; 1999.
[17] Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. Med Decis Making 1992;12:93–108.
[18] van Beek EJR, Kuyer PMM, Schenk BE, Brandjes DPM, ten Cate JW, Büller HR. A normal perfusion lung scan in patients with clinically

suspected pulmonary embolism: frequency and clinical validity. Chest 1995;108:170–3.

[19] van Beek EJR, Kuijer PMM, Büller HR, Brandjes DPM, Bossuyt PMM, ten Cate JW. The clinical course of patients with suspected pulmonary embolism. Arch Intern Med 1997;157:2593–8.

[20] Turkstra F, Kuijer PMM, van Beek EJR, Brandjes DPM, ten Cate JW, Buller HR. Diagnostic utility of ultrasonography of leg veins in patients suspected of having pulmonary embolism. Ann Intern Med 1997;126:775–81.

[21] Miniati M, Prediletto R, Formichi B, Marini C, Di Ricco G, Tonelli L, Allescia G, Pistolesi M. Accuracy of clinical assessment in the diagnosis of pulmonary embolism. Am J Respir Crit Care Med 1999;159:864–71.

[22] Stollberger C, Finsterer J, Lutz W, Stoberl C, Kroiss A, Valentin A, Slany J. Multivariate analyses-based prediction rule for pulmonary embolism. Thromb Res 2000;97:267–73.

[23] Wells PS, Ginsberg JS, Anderson DR, Kearon C, Gent M, Turpie AG, Bormanis J, Weitz J, Chamberlain M, Bowie D, Barnes D, Hirsh J. Use of a clinical model for safe management of patients with suspected pulmonary embolism. Ann Intern Med 1998;129:997–1005.

[24] Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall; 1993.

[25] Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.

[26] Houwelingen van JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303–25.

[27] Banzhaf W, Nordin P, Keller RE, Francone FD. Genetic programming, an introduction. San Fransisco, CA: Morgan Kaufmann Publishers Inc.; 1998.

[28] KiQ Ltd. http://www.kiq.com.