ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 404−412

Journal of Clinical Epidemiology

Development and validation of clinical prediction models: Marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming

Kristel J.M. Janssen[a,*], Ivar Siccama[b], Yvonne Vergouwe[a], Hendrik Koffijberg[a], T.P.A. Debray[a], Maarten Keijzer[c], Diederick E. Grobbee[a], Karel G.M. Moons[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 AB Utrecht, The Netherlands*
[b]*Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands*
[c]*Pegasystems Benelux, Amsterdam, The Netherlands*

Accepted 9 August 2011; Published online 02 January 2012

Abstract

Objective: Many prediction models are developed by multivariable logistic regression. However, there are several alternative methods to develop prediction models. We compared the accuracy of a model that predicts the presence of deep venous thrombosis (DVT) when developed by four different methods.

Study Design and Setting: We used the data of 2,086 primary care patients suspected of DVT, which included 21 candidate predictors. The cohort was split into a derivation set (1,668 patients, 329 with DVT) and a validation set (418 patients, 86 with DVT). Also, 100 cross-validations were conducted in the full cohort. The models were developed by logistic regression, logistic regression with shrinkage by bootstrapping techniques, logistic regression with shrinkage by penalized maximum likelihood estimation, and genetic programming. The accuracy of the models was tested by assessing discrimination and calibration.

Results: There were only marginal differences in the discrimination and calibration of the models in the validation set and cross-validations.

Conclusion: The accuracy measures of the models developed by the four different methods were only slightly different, and the 95% confidence intervals were mostly overlapped. We have shown that models with good predictive accuracy are most likely developed by sensible modeling strategies rather than by complex development methods.

Keywords: Prediction model; Logistic regression; Penalized maximum likelihood estimation; Genetic programming

1. Introduction

In clinical prediction research, patient characteristics, test results, and disease characteristics are often combined in prediction models to estimate the risk that a disease is present (diagnosis) or will occur (prognosis) [1,2]. Model development is a complex process and includes the selection and coding of predictors, testing the need for transformations of continuous predictors, and assessing the predictive strength of predictors. A key warning in model development is to be aware of the risk of overfitting. As a result of overfitting, the predicted risks will be too extreme in future patients; that is, low risks are too low and high risks are too high. This risk of overfitting particularly occurs in relatively small data sets where too many predictors relative to the number of events are studied [3−6]. A rule of thumb is that at least 10 events are needed to consider one candidate predictor (with one degree of freedom) [5,7].

Many prediction models are developed with multivariable regression analysis [2,5,8]. Overfitting of prediction models can be reduced by a priori limiting the number of predictors and correcting the regression coefficients of the predictors using a so-called shrinkage factor [3−6]. This can be a heuristic shrinkage factor, which is based on the fitted model chi-square (in case of regression) corrected for the degrees of freedom that were considered [4]. Also, bootstrapping methods can be used to estimate a shrinkage factor [5,9]. In both methods, a uniform shrinkage factor is estimated post hoc, which can be used to adjust all regression coefficients. Penalized maximum likelihood estimation (PMLE) resembles logistic regression, except that the

Disclosure: At the time of this study, Drs Keijzer and Siccama were affiliated with the company that developed the genetic programming software utilized in the study.

* Corresponding author. Tel.: +31-30-2509380.
E-mail address: kristel.janssen@mapigroup.com (K.J.M. Janssen).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.08.011

regression coefficients are shrunk individually (nonuniform) and also directly during modeling [10–12]. This nonuniform shrinkage is the major advantage of PMLE.

Other methods to develop prediction models are, for example, classification and regression trees and neural networks. However, in medical data, it has been shown that these types of prediction models often do not achieve higher predictive accuracy [13–18]. Genetic programming, however, is a more novel and promising search method that may improve the selection and transformation of predictors, and it may lead to models with good predictive accuracy in new patients [19–22]. The modeling process starts with a large number of candidate prediction models that are stepwise optimized by selecting the best models and adding random variations (see also the Methods section). Although a promising technique, critics state that genetic programming is more prone to overfitting compared with conventional development methods.

Only a few studies have compared models developed by conventional logistic regression vs. models developed by PMLE [3,12] or by genetic programming [19], and direct comparisons are lacking. Using empirical data from a study on the diagnosis of deep venous thrombosis (DVT), we compared the results of four different prediction modeling methods, namely logistic regression analysis without shrinkage, logistic regression analysis with uniform shrinkage, PMLE (nonuniform shrinkage), and genetic programming (no shrinkage). To adhere to daily practice in which methodological guidelines are followed, we explicitly chose a sensible modeling strategy, in which precautions to prevent overfitting were taken. All methods used a derivation set to develop the models and a validation set to test the accuracy of the models.

2. Materials and methods

2.1. Data

Our clinical example concerns the prediction of the presence or absence of DVT. Timely diagnosis of DVT is important because patients with untreated DVT may develop pulmonary embolism, whereas unjustified therapy with anticoagulants poses a risk for major bleeding [23]. In primary care, the physician decides based on patient history, physical examination, and usually the D-dimer value, which patients should be referred to the hospital and which can be safely kept under their own surveillance. A diagnostic prediction model can aid physicians in this decision.

We used the data of a cohort of 2,086 primary care patients with a suspicion of DVT that were included between 2001 and 2005, described previously in the literature [24,25]. After patient history, physical examination, and the D-dimer value were obtained (in total 21 candidate predictors, see Table 1), all patients underwent repeated leg ultrasound as the reference method to determine the true presence or absence of DVT. To compare the effect of the different modeling strategies, we randomly split the cohort into a derivation set (80% of the data: 1,668 patients, 329 with DVT) to develop the models and a validation set (418 patients, 86 with DVT) to test the models (Table 1).

Table 1. Distribution of the 21 predictors and the outcome in the derivation set and the validation set, *n* (%) unless stated otherwise

Patient characteristics	Derivation set ($n = 1,668$)	Validation set ($n = 418$)
Age (yr)[a]	60 (18)	61 (17)
Male gender, *n* (%)	603 (36)	165 (40)
Oral contraception use, *n* (%)	172 (10)	32 (8)
Hormonal replacement use, *n* (%)	32 (2)	8 (2)
Duration of symptoms (d)[a]	8 (9)	8 (7)
Absence of a leg trauma, *n* (%)	1,402 (84)	353 (84)
Previous DVT, *n* (%)	343 (20)	78 (19)
Family history of DVT, *n* (%)	341 (20)	96 (23)
Presence of a malignancy, *n* (%)	95 (6)	19 (5)
Immobilization, *n* (%)	217 (13)	61 (15)
Recent surgery, *n* (%)	198 (12)	62 (15)
Swelling whole leg, *n* (%)	727 (44)	205 (49)
Vein distension, *n* (%)	300 (18)	70 (17)
Pain in leg, *n* (%)	1,458 (87)	347 (83)
Pain when walking, *n* (%)	1,370 (92)	324 (78)
Oedema in leg, *n* (%)	1,039 (62)	271 (65)
Ill feeling, *n* (%)	315 (19)	85 (20)
Tender venous system, *n* (%)	1,213 (73)	299 (72)
Pregnancy, *n* (%)	38 (2)	8 (2)
Log (calf circumference)[a]	1.07 (0.52)	1.09 (0.51)
Log (D-dimer level)[a]	6.81 (1.16)	6.89 (1.14)
DVT present, *n* (%)	329 (20)	86 (20)

Abbreviation: DVT, deep venous thrombosis.
[a] Mean (standard deviation).

2.2. Methods to develop the prediction models in the derivation set

2.2.1. Logistic regression without shrinkage

Logistic regression was used to assess the association between each continuous variable (age, duration of pain, difference in calf circumstance, and the D-dimer value) with the log odds of the presence of DVT. For each continuous variable, restricted cubic splines with three knots were used to model the relationship between the variable and the presence of DVT [5,6]. If the plot of the restricted cubic spline showed that the association between the continuous predictors and the outcome was not linear, the predictors were transformed accordingly. Subsequently, all 21 candidate predictors were included in a logistic regression model. Backward stepwise elimination of the candidate predictors was applied to fit the final prediction model. To eliminate predictors from the model, we used the Akaike information criterion. This implies that the increase in χ^2 has to be larger than two times the degrees of freedom. When considering eliminating a predictor with one degree of freedom, this corresponds to a P-value >0.157.

2.2.2. Logistic regression with one uniform shrinkage factor obtained by bootstrapping

The same modeling steps were followed as in the previous method. Subsequently, to obtain the uniform shrinkage factor, bootstrapping techniques were applied [5,6]. We drew 101 bootstrap samples from the derivation set. In each bootstrap sample, the modeling process was repeated, including testing for transformations and the backward stepwise elimination of the candidate predictors. This resulted in 100 models that were applied to the original derivation set. The uniform shrinkage factor was the mean of the 100 calibration slopes (see also the description of the calibration plots in the section ''Accuracy measures''). To shrink the regression coefficients of the model, these were multiplied with this shrinkage factor.

2.2.3. Logistic regression with nonuniform shrinkage by PMLE

As in the conventional logistic regression, the linearity of the continuous predictors was studied, and all 21 candidate predictors were included in a logistic regression model. For an extensive description of PMLE, we refer to the Appendix (see on the journal's Web site at www.elsevier.com) and the literature [6,12]. In brief, instead of maximizing the log likelihood as in conventional logistic regression, PMLE maximizes the penalized log likelihood. Hence, the maximum log likelihood of the full model is adjusted (shrunk) by a penalty factor. Accordingly, the estimated regression coefficients are individually (nonuniform) adjusted for overfitting during the model fit. Because of the penalization, the number of degrees of freedom effectively used in PMLE is lower than the actual number of predictors, reducing the potential for overfitting [6,10,26].

As there is no selection of candidate predictors, a prediction model that has been developed by PMLE includes all possible candidate predictors, which can be impractical to use in clinical practice. Alternatively, a parsimonious model can be obtained by estimating a new model with a reduced number of predictors. This parsimonious model approximates the predictions from the penalized model [6]. To develop such parsimonious model, ordinary least squares (linear) regression is used to fit the linear predictor of the penalized model as the outcome and all 21 predictors as covariates [6]. This model necessarily has an R^2 of 1. Next, backward stepwise elimination is used to exclude the least important predictors until the R^2 is lower than 0.975. The regression coefficients of the predictors that remained in this linear regression model are also penalized for overfitting as the shrinkage that is used in the penalized model is inherited by the reduced model [6]. For a more profound description of this procedure, we refer to the literature [6].

2.2.4. Genetic programming

Genetic programming is a search method for the optimal solutions for a given problem (in this situation, the optimal prediction model), inspired by the biological model of evolution [20,27−31]. For a detailed description of the method, we refer to the Appendix (see on the journal's Web site at www.elsevier.com). In brief, first a set of 50 different prediction models was randomly created, in which the prediction models were different mathematical formulas using different predictors. This can be seen as the first round. Then, in an iterative process, models with a large receiver operating characteristic (ROC) have a higher probability of being selected for the next round. In each round, given two randomly selected models, crossover is realized by randomly swapping parts of the model. In addition, mutations occur by exchanging part of the model with a randomly created substitute. The performance of the models is estimated, and models with a large ROC have a higher probability of being selected for the next round. Consequently, the process of crossover and mutation is repeated in the models that have been selected for the next round. This iterative process was terminated when no significant model improvement was observed, and the model with the largest ROC area was selected as the final genetic programming model.

A prediction model developed by genetic programming can be represented as a binary tree (see Fig. 1). To limit the risk of overfitting, the trees were restricted to be no more than four levels deep, corresponding to a maximum of eight predictors. The building blocks of the model are mathematical operators, chosen from a library of 20 operators (e.g., x, x^2, and $\sin(\pi \times x)$; see also Fig. 1). Each operator has two input values and one output value (Fig. 1). The output of the complete formula for individual patients is a score. To transform this score to a risk of DVT presence, we estimated a logistic regression model in the derivation set. The score is the only covariate, modeled with a restricted cubic spline function, as it most likely has a nonlinear association with the outcome [6]. Subsequently, this model was used to calculate the risk of DVT for each patient from the corresponding score (see the equations provided in the caption of Fig. 1). Note that the genetic programming model cannot be adjusted for overfitting by shrinkage as there are no regression coefficients.

For the present analyses, we used the Predictive Analytics Director (Chordiant Software Inc., Amsterdam, The Netherlands; www.chordiant.com).

2.3. Accuracy of the four models in the validation set

We estimated the accuracy of the models by assessing the discrimination and calibration. Discrimination is the ability of a prediction model to distinguish between patients with the outcome and patients without the outcome. It can be quantified with the area under the ROC curve (ROC area) [32]. An ROC area ranges from 0.5 (no discrimination, same as flipping a coin) to 1.0 (perfect discrimination) [33].

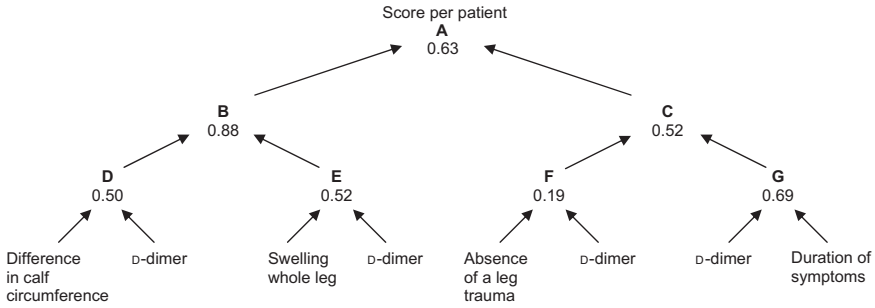

Fig. 1. The final model created by genetic programming presented as a binary tree. The nodes A–G represent the binary operators described below, in which the parameters x (left arrow) and y (right arrow) are the inputs of each operator, and "p" refers to the observed proportion of the outcome (deep venous thrombosis prevalence). The numbers beneath each letter express the weight that this node receives in the calculation of the score for each individual patient.

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

$$B = p\left(2x - k\left((2x-1)^3\right) - 0.5\right) + (1-p)\left(2y - k\left((2y-1)^3\right) - 0.5\right) \text{ and } k = 0.593$$

$$C = 1 - \left(p\sqrt{(1-x^2)}\right) - (1-p)\sqrt{(1-y^2)}$$

$$D = p\left((x-1)^3 + 1\right) + (1-p)\sin(\pi * y)$$

$$E = p\sqrt{x} + (1-p)y^2$$

$$F = px + (1-p)y$$

$$G = px^2 + (1-p)y^2$$

The predicted probabilities can be calculated by $\left(\frac{1}{1+e^{-\text{linear predictor}}}\right)$, in which linear predictor $= -28.79 + 37.74 \times \text{score} - 344.41 \times (\text{score} - 0.62)_+^{\,3} + 489.40 \times (\text{score} - 0.70)_+^{\,3} - 144.99 \times (\text{score} - 0.90)_+^{\,3}$.
Notation $(x)_+$ means $(x)_+ = x$ if $x > 0$ and 0 otherwise.

Calibration refers to the agreement between the predicted risks and observed frequencies of the outcome. It can be assessed with a calibration plot with the predicted risks on the x-axis and the observed frequencies on the y-axis, that is, a smoothed curve of grouped proportions vs. mean predicted probability in groups (see Fig. 2) [8]. The calibration plot shows a calibration line, which can be described by a calibration slope and intercept. These are estimated by fitting the linear predictor of the prediction model as the only covariate in a logistic regression model. The calibration slope of a model in new patients is ideally equal to 1, implying that the calibration plot lies exactly on the 45° line. The calibration intercept of a model in new patients is ideally equal to 0, implying that the mean predicted probability is equal to the mean observed frequency, and the calibration plot crosses the y-axis in 0. A slope <1 indicates overfitting (predicted risks are too extreme), whereas a slope >1 indicates that the predicted risks are not extreme enough. When the slope is not equal to 1, the interpretation of the intercept is less straightforward. Hence, we estimated the intercept with the slope fixed at 1 (calibration in the large) [8]. When this intercept is close to 0, the so-called calibration in the large is good, that is, the mean predicted risk equals the mean observed frequency.

2.4. Accuracy of the four models in cross-validations

To mimic as much as possible current practice in the domain of prediction research, we explicitly chose to conduct our analysis in a particular (usually larger) derivation study sample and a second (usually smaller) validation study, as this is often how prediction models are developed and validated. Yet, the predictive accuracy of any developed model can be influenced by chance or sampling variation [5,34]. Therefore, to quantify the robustness of the developed models (by each method), we also conducted 100 cross-validations in the full cohort to assess the accuracy of the developed models in different samples. The derivation set was 10 times randomly split in 10 stratified equally sized samples. These 10 equally sized samples were combined 10 times, and this approach was repeated 10 times, so that in 100 models were developed on 90% of the data, and tested in the other 10% of the data. This way, 100 cross-validations were thus performed.

With these approaches, we aimed not only to mimic current prediction research as much as possible but also to adjust for potential chance findings. In the cross-validation, the median value of the discrimination and calibration was estimated. Also, paired t-tests were used

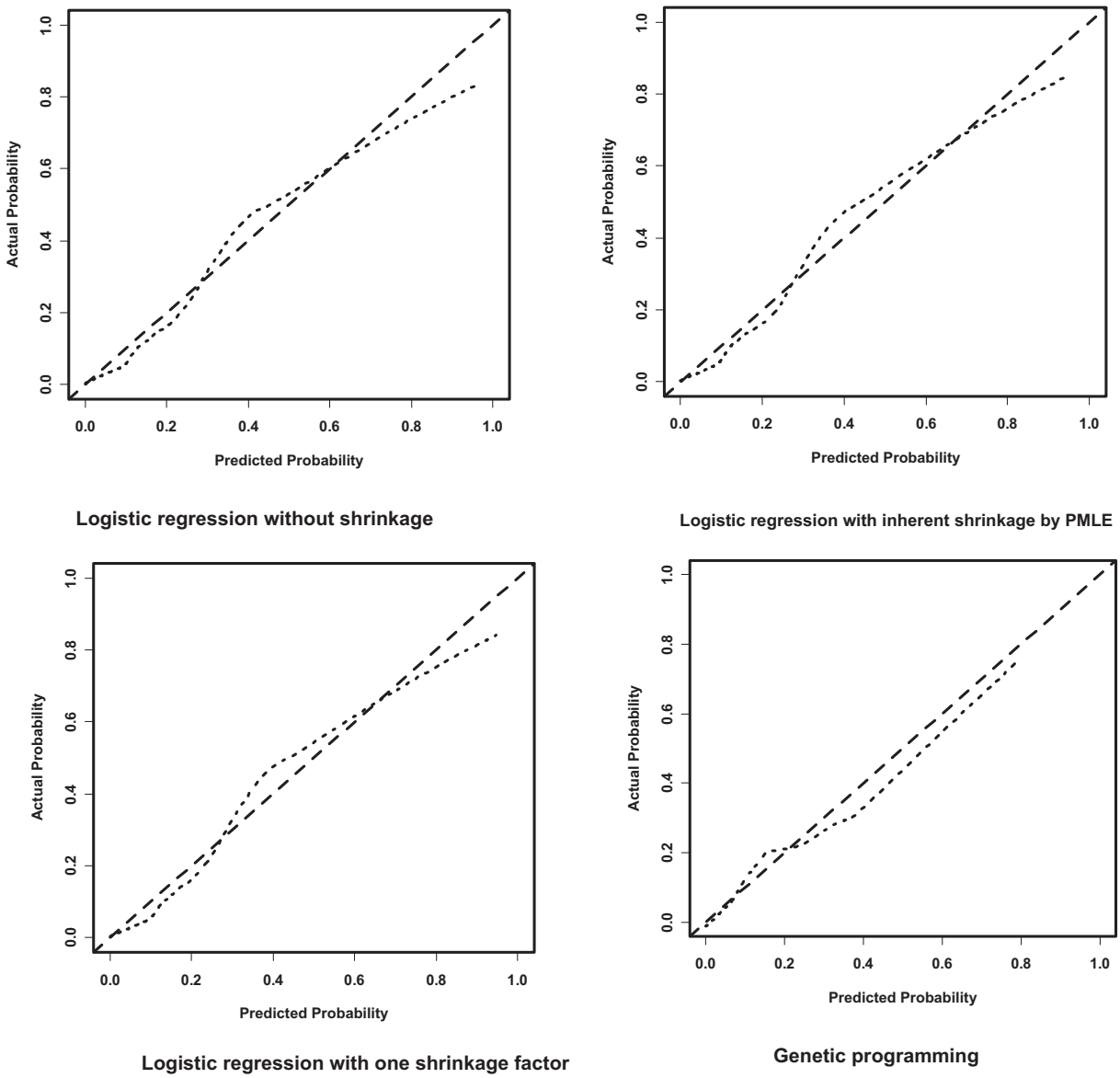

Fig. 2. Calibration plots in the validation set of the models derived by the four derivation methods. PMLE, penalized maximum likelihood estimation.

to assess whether there were significant differences in the discrimination and calibration between the 100 cross-validations.

3. Results

Table 1 shows the distribution of the candidate predictors in the derivation and validation set. The incidence of DVT was similar in both sets (20%). The validation set contained more males compared with the derivation set (40% vs. 36%). More patients in the validation set experienced a swelling of the whole leg (49% vs. 44%), and less patients in the validation set experienced pain when walking (78% vs. 92%).

3.1. Model derivation

3.1.1. Logistic regression without shrinkage

The difference in calf circumference and the D-dimer value showed a logarithmic association with the log odds of the presence of DVT and were subsequently transformed. The final prediction model included six predictors: age, duration of symptoms, the absence of a leg trauma, pregnancy, the difference in calf circumference, and the D-dimer value (Table 2). The ROC area in the derivation set was 0.904 (95% confidence interval [CI]: 0.885–0.922).

3.1.2. Logistic regression with one uniform shrinkage factor

By definition, the final prediction model included the same predictors as the previous model. Bootstrapping

Table 2. Regression coefficients of the models derived by logistic regression: without shrinkage, with uniform shrinkage by bootstrapping, and with nonuniform shrinkage by PMLE

	Conventional logistic regression without shrinkage	Conventional logistic regression with shrinkage	PMLE
Intercept	−15.871	−14.970	−15.562
Age	−0.024	−0.023	−0.021
Duration of symptoms	−0.029	−0.027	−0.027
Absence of leg trauma	0.688	0.647	0.625
Pregnancy	−1.764	−1.658	NA
Log (difference in calf circumference)	1.010	0.949	1.019
Log (D-dimer value)	1.967	1.849	1.892

Abbreviations: NA, not applicable; PMLE, penalized maximum likelihood estimation.

The formula of the above described prediction models takes the form: linear predictor $= \beta_0 + \beta_1 \times \text{predictor}_1 + \beta_2 \times \text{predictor}_2 + \ldots + \beta_n \times \text{predictor}_n$, where β_0 is the intercept and β_1 till β_n are the regression coefficients of the predictors. The probability of deep venous thrombosis in an individual patient (scale: 0–100%) can be calculated by $\frac{1}{1+e^{-\text{linear predictor}}}$.

resulted in a shrinkage factor of 0.94 that was used to adjust the regression coefficients (Table 2). The ROC area was 0.904 (95% CI: 0.885–0.922).

3.1.3. Logistic regression with nonuniform shrinkage by PMLE

The reduced penalized model included five predictors; the same predictors as in the previous model although without pregnancy (Table 2). The optimal penalty factor was three, and the penalized model had 18.93 effective degrees of freedom (vs. 21 degrees of freedom in conventional logistic regression). The ROC area was 0.902 (95% CI: 0.883–0.921).

3.1.4. Genetic programming

The final prediction model included nearly the same five predictors as the model developed by logistic regression with PMLE, except that age was excluded and swelling of the leg was included (Fig. 1). The ROC area was 0.910 (95% CI: 0.893–0.928).

3.2. Accuracy of the four models in the validation set

We found only marginal differences in the discriminative ability of the four models in the validation set (Table 3). All methods led to calibration slopes close to 1 (Fig. 2 and Table 3). The slopes of the models developed by logistic regression without shrinkage and genetic programming were smaller than one, indicating overfitting (Table 3). The intercept (given that the slope was equal to 1) was slightly negative although close to 0 for all models (Fig. 2 and Table 3). PMLE led to the best calibration in the large (intercept = −0.087) and genetic programming to the worst (intercept = −0.160).

3.3. Accuracy of the four models in the cross-validations

Table 4 shows that there were only marginal differences in the median discriminative ability of the models in the 100 cross-validations (logistic regression with and without shrinkage: 0.902, PMLE: 0.901, and genetic programming: 0.883). The paired *t*-test yielded no significant differences in ROC area for the logistic regression models (with and without shrinkage) and PMLE. The ROC areas of the genetic programming models were significantly different from the ROC areas of the logistic regression models (with and without shrinkage) ($P = 0.04$) and PMLE ($P = 0.05$).

All methods led to models with a median calibration slope close to 1. The paired *t*-tests showed statistically significant differences between the logistic regression models with and without shrinkage ($P = 0.01$, slope 0.992 and 0.980, respectively), between the logistic regression models with shrinkage and genetic programming models ($P < 0.001$, slope 0.992 and 0.945, respectively), and between the PMLE and genetic programming models ($P < 0.001$, slope 1.009 and 0.945, respectively). The median intercept (given slope = 1) was close to 0 for all models and not significantly different between the different methods.

4. Discussion

We used four different methods to develop a prediction model to predict the presence or absence of DVT: logistic regression without shrinkage, logistic regression with one uniform shrinkage factor (estimated from bootstrapping

Table 3. Comparison of the accuracy measures of the models using the four different derivation methods in the validation set

Method	Logistic regression	Logistic regression	PMLE	Genetic programming
Shrinkage	—	Uniform shrinkage	Nonuniform shrinkage	—
ROC area (95% CI)	0.906 (0.872–0.941)	0.906 (0.872–0.941)	0.907 (0.873–0.941)	0.912 (0.882–0.943)
Calibration slope (95% CI)	0.961 (0.756–1.169)	1.024 (0.804–1.244)	1.027 (0.808–1.247)	0.982 (0.788–1.176)
Calibration intercept given slope = 1 (95% CI)	−0.142 (−0.455, 0.171)	−0.105 (−0.411, 0.201)	−0.087 (−0.394, 0.220)	−0.160 (−0.480, 0.161)

Abbreviations: PMLE, penalized maximum likelihood estimation; ROC, receiver operating characteristic; 95% CI, 95% confidence interval.

The accuracy measures include discrimination, expressed by the ROC area, and calibration, expressed by the slope and the intercept/slope = 1 (calibration in the large).

Table 4. Comparison of the predictive accuracy of the four differently derived models averaged across the 100 cross-validations

Method	Logistic regression	Logistic regression	PMLE	Genetic programming
Shrinkage	—	Uniform shrinkage	Nonuniform shrinkage	—
ROC area (25–75%)	0.902 (0.884–0.917)	0.902 (0.884–0.917)	0.901 (0.882–0.917)	0.883 (0.864–0.902)
Calibration slope (25–75%)	0.980 (0.865–1.112)	0.992 (0.879–1.128)	1.009 (0.895–1.149)	0.945 (0.860–1.026)
Calibration intercept given slope = 1 (25–75%)	−0.022 (−0.160, 0.123)	−0.011 (−0.148, 0.136)	0.031 (−0.106, 0.181)	−0.063 (−0.186, 0.079)

Abbreviations: PMLE, penalized maximum likelihood estimation; ROC, receiver operating characteristic.

The accuracy measures (median with 25% and 75% quantiles) include discrimination, expressed by the ROC area, and calibration, expressed by the calibration slope and the calibration intercept/slope = 1 (calibration in the large). 25–75%: 25–75% quantiles.

techniques), logistic regression with nonuniform shrinkage by PMLE, and genetic programming. The accuracy of the methods was tested in 100 cross-validations and new patients (external validation). There were only marginal differences in the discriminative ability of the models in the cross-validations. Similarly, the accuracy measures of the four models in the validation set were only slightly different, and the 95% CIs were largely overlapped.

We expected that the PMLE models would show better discriminative ability (because of less overfitting) in the validation set than the logistic regression models. This was not shown in our data. In addition, we expected that logistic regression without shrinkage and genetic programming would lead to the most overfitted models (calibration slopes smaller than 1). Although this was indeed shown in the validation set, it was only partly shown in the cross-validations, where only genetic programming resulted in (slightly) overfitted models. There are several possible reasons why the amount of overfitting was so low.

First, we had a relatively large data set while the risk of overfitting is higher in small data sets [3–6]. Our derivation set consisted of 1,668 patients, of which 329 (20%) had DVT. According to the rule of thumb that at least 10 events are needed to consider one candidate predictor (with one degree of freedom), this implies that 33 candidate predictors could have been considered. We considered only 21 candidate predictors, using eight extra degrees of freedom for assessing the linearity of the four continuous predictors with restricted cubic splines. Hence, within total 29°, we still conformed to that rule of thumb, reducing the risk of overfitting.

Second, to adhere to daily practice in which methodological guidelines are followed, we explicitly chose a sensible modeling strategy, in which precautions to prevent overfitting were taken. Because of its flexibility, genetic programming models are more prone to overfitting, as besides the ''true'' associations also ''noise'' associations may be modeled. However, a sensible application of genetic programming reduces the risk of overfitting. We conducted a simple experiment in which basic safeguards for both logistic regression and genetic programming were removed. For genetic programming, the predictors were not grouped based on statistical significance of predictive behavior, and the predictors were modeled nonlinear. For logistic regression, all continuous predictors were modeled with a restricted cubic spline. We also included interaction terms between the D-dimer value and all other predictors. We did not shrink the regression coefficients. The discrimination of the genetic programming model decreased from 0.94 in the derivation set to 0.88 in the validation set. For the logistic regression model, it slightly decreased from 0.92 to 0.90. However, the calibration slope of the logistic regression model was equal to 0.70, indicating serious overfitting. Hence, although some modeling methods are more prone to overfitting, the actual extent of overfitting is largely determined by thoughtfulness of the modeling process.

Third, logistic regression and PMLE is based on maximizing the (penalized) log likelihood, whereas genetic programming maximizes the c-statistic (area under the ROC curve). Because the c-statistic is not the most sensitive measure to assess incremental value of predictors [35,36], this may have decreased the risk of overfitting for the genetic programming model.

All four methods have their own opportunities and pitfalls, depending on the characteristics of the data and situation at hand. The results of our study, in combination to existing knowledge from previous studies, lead to the following advices. First, models with good predictive accuracy in new patients were most likely developed by sensible modeling strategies rather than by complex development methods. Second, when conducting logistic regression, one should preferably apply shrinkage, especially in small data sets [5,6,37,38]. Although the discriminative ability of a model will not be changed, in general the calibration of the model in future patients will be improved. Third, when evaluating the performance of prediction models, not only the discrimination but also the calibration should be assessed [35]. Fourth, when many interaction terms are considered, PMLE should be preferred over shrinkage by one uniform shrinkage factor, especially when the data set is (relatively) small. In our data, we had no clinical reasons to include interaction terms. Therefore, the effectively used number of degrees of freedom in PMLE was only slightly lower than the actual number of predictors (18.93 vs. 21). Fifth, genetic programming is highly flexible in the transformations of continuous predictors. Especially in data sets with many continuous predictors, genetic programming can provide insight in unconventional associations between the predictor and the outcome. However, in our data set, only four of the 21 candidate predictors were continuous predictors. Note that the use of

restricted cubic spline functions and fractional polynomials [39−41] in logistic regression also increases the flexibility of modeling continuous predictors. Sixth, genetic programming is developed as a flexible search strategy, that is, the optimal fit is searched in a data set. Therefore, it can be used in data sets where no predefined clinical associations exist. It may be a promising tool to find nonlinear associations and unknown interactions between predictors. Seventh, logistic regression and PMLE modeling is based on maximizing the (penalized) log likelihood, whereas genetic programming can be used to maximize different parameters or several parameters at the same time. For example, in cost-effectiveness analyses, one wants to optimize both utilities and costs at the same time. Genetic programming can be applied to find the model that best optimizes both aspects. Eighth, the models developed by genetic programming cannot be shrunk according to standard methodology. Yet, by the method presented in this article, the scores that are calculated with the model can be recalibrated in such a way that the predicted risks are equal to the observed frequencies.

In conclusion, we found no major differences in the predictive accuracy of the models, and the 95% CIs of the accuracy measures mostly overlapped. In a data set with one strong predictor that shows a linear association with the outcome or an association that can be easily approximated by a restricted cubic spline or transformation (as the D-dimer in our study), modeling methods other than logistic regression will probably have similar or only slightly better accuracy than logistic regression. The choice between the development methods should be based on the characteristics of the data and situation at hand. In our case, models with good predictive accuracy were most likely developed by sensible modeling strategies rather than by complex development methods.

Acknowledgments

The authors gratefully acknowledge the financial contribution by the Netherlands Orgsanisation for Scientific Research (ZonMw project 918.10.615 and 016.046.360).

Appendix

Supplementary material

Supplementary material can be found, in the online version, at doi:10.1016/j.jclinepi.2011.08.011.

References

[1] Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201−10.
[2] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488−94.
[3] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19:1059−79.
[4] Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303−25.
[5] 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;15:361−87.
[6] Harrell FE Jr. Regression modelling strategies. New York, NY: Springer-Verlag; 2001.
[7] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373−9.
[8] Steyerberg EW. Clinical prediction models. A practical approach to development, validation, and updating. Springer; 2009.
[9] Efron B. Censored data and the bootstrap. J Am Stat Assoc 1981;76:312−9.
[10] Gray RJ. Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 1997;87:942−51.
[11] Verweij PJ, van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med 1994;13:2427−36.
[12] Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 2004;57:1262−70.
[13] 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.
[14] 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.
[15] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49:1225−31.
[16] 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.
[17] 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.
[18] Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med 2007;26:2937−57.
[19] Biesheuvel CJ, Siccama I, Grobbee DE, Moons KG. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. J Clin Epidemiol 2004;57:551−60.
[20] Forrest S. Genetic algorithms: principles of natural selection applied to computation. Science 1993;261:872−8.
[21] Podbregar M, Kovacic M, Podbregar-Mars A, Brezocnik M. Predicting defibrillation success by 'genetic' programming in patients with out-of-hospital cardiac arrest. Resuscitation 2003;57:153−9.
[22] Tsakonas A, Dounias G, Jantzen J, Axer H, Bjerregaard B, von Keyserlingk DG. Evolving rule-based systems in two medical domains using genetic programming. Artif Intell Med 2004;32:195−216.
[23] Hirsh J, Hoak J. Management of deep vein thrombosis and pulmonary embolism. A statement for healthcare professionals. Council on Thrombosis (in consultation with the Council on Cardiovascular Radiology), American Heart Association. Circulation 1996;93:2212−45.
[24] Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005;94:200−5.
[25] Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KG. Excluding deep vein thrombosis safely in primary care. J Fam Pract 2006;55:613−8.

[26] Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. Stat Med 2002;21: 3803−22.
[27] Barrett J, Kostadinova A, Raga JA. Mining parasite data using genetic programming. Trends Parasitol 2005;21:207−9.
[28] Koza JR. Genetic programming III. Cambridge, MA: MIT Press; 1999.
[29] Poli R, McPhee NF. General schema theory for genetic programming with subtree-swapping crossover: part I. Evol Comput 2003;11:53−66.
[30] Holland JH. Adaptation in natural and artificial systems. Ann Arbor, MI: University of Michigan Press; 1975.
[31] Goldberg DE. Genetic algorithms in search optimization and machine learning. Addison Wesley Publishing Company; 1989.
[32] Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143−52.
[33] Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839−43.
[34] Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York, NY: Chapman & Hall; 1993.
[35] Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928−35.
[36] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157−72.
[37] 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.
[38] Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56:441−7.
[39] Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999;28:964−74.
[40] Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. Stat Med 2003;22:639−59.
[41] Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 2004;23:2509−25.