STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2937–2957
Published online 21 December 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2770

A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality

Peter C. Austin[1, 2, 3, *, †]

[1] *Institute for Clinical Evaluative Sciences, Toronto, Ont., Canada*
[2] *Department of Public Health Sciences, University of Toronto, Canada*
[3] *Department of Health Management, Policy and Evaluation, University of Toronto, Canada*

SUMMARY

Clinicians and health service researchers are frequently interested in predicting patient-specific probabilities of adverse events (e.g. death, disease recurrence, post-operative complications, hospital readmission). There is an increasing interest in the use of classification and regression trees (CART) for predicting outcomes in clinical studies. We compared the predictive accuracy of logistic regression with that of regression trees for predicting mortality after hospitalization with an acute myocardial infarction (AMI). We also examined the predictive ability of two other types of data-driven models: generalized additive models (GAMs) and multivariate adaptive regression splines (MARS). We used data on 9484 patients admitted to hospital with an AMI in Ontario. We used repeated split-sample validation: the data were randomly divided into derivation and validation samples. Predictive models were estimated using the derivation sample and the predictive accuracy of the resultant model was assessed using the area under the receiver operating characteristic (ROC) curve in the validation sample. This process was repeated 1000 times—the initial data set was randomly divided into derivation and validation samples 1000 times, and the predictive accuracy of each method was assessed each time. The mean ROC curve area for the regression tree models in the 1000 derivation samples was 0.762, while the mean ROC curve area of a simple logistic regression model was 0.845. The mean ROC curve areas for the other methods ranged from a low of 0.831 to a high of 0.851. Our study shows that regression trees do not perform as well as logistic regression for predicting mortality following AMI. However, the logistic regression model had performance comparable to that of more flexible, data-driven models such as GAMs and MARS. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: logistic regression; regression trees; classification trees; predictive model; validation; recursive partitioning; generalized additive models; multivariate adaptive regression splines; acute myocardial infarction

[*] Correspondence to: Peter C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ont., Canada M4N 3M5.
[†] E-mail: peter.austin@ices.on.ca

Contract/grant sponsor: Ontario Ministry of Health and Long Term Care
Contract/grant sponsor: Canadian Institutes of Health Research (CIHR)

Received 28 November 2005
Accepted 19 October 2006

1. INTRODUCTION

There is an increasing interest in predicting the probability of adverse events for patients hospitalized for medical or surgical treatment. Accurately predicting the probability of adverse events allows for effective patient risk stratification, thus permitting more appropriate medical care to be delivered to patients [1–4]. Furthermore, accurately predicting the probability of an adverse event allows for risk-adjusted outcomes to be compared across providers of health care [5].

Logistic regression is the most commonly used method for predicting the probability of an adverse outcome in the medical literature. Recently, data-driven methods, such as classification and regression trees (CART) have been used to identify subjects at increased risk of adverse outcomes or of increased risk of having specific diagnoses [6–41]. Advocates for CART have suggested that these methods allow the construction of easily interpretable decision rules that can easily be applied in clinical practice. Furthermore, CART methods are adept at identifying important interactions in the data [31, 34, 40] and in identifying clinical subgroups of subjects at very high or very low risk of adverse outcomes [41].

Several studies have compared the performance of regression trees and logistic regression for predicting outcomes. These studies can be grouped into three broad categories. First, studies that compared the variables identified by logistic regression as significant predictors of the outcome with those variables identified by a regression tree analysis as predictors of the outcome [6–16]. Second, studies that compared the sensitivity and specificity of logistic regression with that of regression trees [6, 12, 17–30]. Third, a small number of studies that compared the predictive accuracy, as measured by the area under the receiver operating characteristic (ROC) curve, of logistic regression with that of regression trees [13, 14, 31–39, 42]. The first category of studies does not allow one to compare the predictive ability of the two different prediction methods. Rather, it compares agreement on which factors are prognostically important. Since each model uses variables in a different manner, it is possible that the methods could differ in predictive accuracy, yet agree on which factors are prognostically important. The second category of studies compares sensitivity and specificity of regression trees with that of logistic regression. However, computing sensitivity and specificity from a logistic regression model requires specifying a probability threshold, and then assuming that the response will be positive if the predicted probability exceeds this probability threshold. Harrell criticizes this approach for several reasons [43]. In particular, it is highly dependent upon the probability threshold chosen for a positive prediction. Furthermore, it is an insensitive and inefficient measure of predictive accuracy [44].

Only a small number of studies have compared the predictive ability of regression trees with that of logistic regression using the area under the ROC curve [13, 14, 31–39, 42]. Among these studies, the conclusions were inconsistent. Six studies concluded that regression trees and logistic regression had comparable performance [13, 31, 33, 36–38]; five studies concluded that logistic regression had superior performance to regression trees [14, 32, 34, 39, 42]; while one study arrived at the opposite conclusion [35]. Only one recent study, using a relatively small sample, employed repeated split sample validation to examine the robustness of the findings to the particular splitting of the sample in derivation and validation samples [38]. The authors of this study suggested that similar methods be applied in other disciplines and other data sets to test the validity of their findings [38].

The current study had two objectives. First, to compare the predictive ability of conventional logistic regression with that of regression tree methods for predicting 30-day mortality in a sample of patients admitted to hospital with a diagnosis of acute myocardial infarction (AMI or heart attack). Second, to compare the relative performance of two other data-driven methods of analysis:

DOI: 10.1002/sim

generalized additive models (GAMs) and multivariate adaptive regression spline (MARS) models. We used repeated split sample validation using a large sample of patients hospitalized with AMI.

2. METHODS

2.1. Data sources

Detailed clinical data were available on a sample of 9484 patients discharged from 102 Ontario hospitals between 1 April 1999 and 31 March 2001. Data were obtained by retrospective chart review. These data were collected as part of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, an ongoing initiative intended to improve the quality of care for patients with cardiovascular disease in Ontario [45]. Data on patient history, cardiac risk factors, comorbid conditions and vascular history, vital signs, and laboratory tests were collected for this sample. Prevalence of dichotomous variables and medians and the 25th and 75th percentiles of continuous variables considered in the current study are reported in Table I. Patient records were linked to the Registered Persons Database (RPDB) using encrypted health card numbers that allowed us to determine the vital status of each patient at 30-day following admission. Overall, 1065 (11.2 per cent) patients died within 30-day of admission.

2.2. Initial model for predicting 30-day AMI mortality

In an earlier study, a model-selection method based on repeated bootstrap resampling was used to develop a parsimonious logistic regression model for predicting AMI mortality. The resultant model consisted of the following terms: age, presence of cardiogenic shock at admission, systolic blood pressure at admission, family history of coronary artery disease (CAD), respiratory rate at admission, glucose, white blood count, and creatinine [46]. This model was developed by drawing repeated bootstrap samples from a sample of AMI patients. Models predicting AMI mortality were developed using backwards elimination on each bootstrap sample. Those variables that were identified as significant predictors of AMI mortality in at least 60 per cent of the bootstrap samples were retained for inclusion in the final predictive model. This method of model derivation has been shown to result in predictive performance similar to using cross-validation or Akaike's information criterion (AIC) for model selection [46]. Family history of CAD and presence of cardiogenic shock are dichotomous variables. The remaining six variables are continuous variables and were entered linearly in the regression model. This model will be used as the basis for some of the regression models that we will consider in this study.

2.3. Predictive models for AMI mortality

In this section, we describe the four different classes of predictive models that were used to predict 30-day AMI mortality. All model fitting and model validation was done using the R statistical programming language [47].

2.3.1. Logistic regression. Three separate logistic regression models were used to predict 30-day AMI mortality. The first model consisted of the eight variables described above: age, presence of cardiogenic shock at admission, systolic blood pressure at admission, family history of CAD, respiratory rate at admission, glucose, white blood count, and creatinine [46]. The second model

DOI: 10.1002/sim

Table I. Available variables and characteristics of the study sample.

Prevalence of dichotomous variables	
Female	36.0 per cent
Presenting signs and symptoms	
Acute congestive heart failure (acute CHF)/pulmonary oedema	5.7 per cent
Cardiogenic shock	1.6 per cent
Classic cardiac risk factors	
Diabetes	26.3 per cent
History of hypertension	46.0 per cent
Smoking history	32.3 per cent
History of cerebrovascular accident or transient ischaemic attack	10.3 per cent
History of hyperlipidaemia	30.6 per cent
Family history of coronary artery disease	30.2 per cent
Comorbid conditions	
Angina	33.0 per cent
Cancer	3.1 per cent
Dementia	3.9 per cent
Peptic ulcer disease	5.5 per cent
Previous acute myocardial infarction	23.1 per cent
Asthma	5.5 per cent
Depression	7.2 per cent
Peripheral arterial disease	7.7 per cent
Previous PCI	3.2 per cent
Congestive heart failure (chronic)	5.0 per cent
Hyperthyroidism	1.3 per cent
Previous CABG surgery	6.7 per cent
Aortic stenosis	1.7 per cent
Medians of continuous variables (25th percentile–75th percentile)	
Age	69 (57–78)
Vital signs on admission	
Systolic blood pressure on admission	146 (126–168)
Diastolic blood pressure on admission	82 (70–95)
Heart rate on admission	81 (68–98)
Respiratory rate on admission	20 (18–23)
Laboratory test results	
Haemoglobin	139 (127–151)
White blood count	9.6 (7.7–12.2)
Sodium levels	139 (137–141)
Potassium levels	4.1 (3.7–4.4)
Glucose levels	7.85 (6.4–10.9)
Urea level	6.5 (5.1–8.7)
Creatinine levels	93 (78–115)

was constructed using backwards variable elimination. The initial model consisted of the above eight main effects and all two-way interactions between these main effects, with the eight main effects being forced to remain in each model. The third model was also constructed using backwards variable elimination. However, in this instance, the initial model consisted of the 34 variables listed in Table I. The logistic regression models were fit using the glm function in R.

Backwards variable elimination was done using the step function in R. This implementation of backwards variable elimination is based upon sequentially eliminating variables from an initial model. At each step the variable is removed from the current model that results in the greatest reduction in the AIC. The process of eliminating variables terminates either when a pre-specified boundary model is achieved or when no step will cause a further reduction in the AIC criterion [48]. For the second model above, the boundary model consisted of eight main effects: age, presence of cardiogenic shock at admission, systolic blood pressure at admission, family history of CAD, respiratory rate at admission, glucose, white blood count, and creatinine: the eight variables in the first logistic regression model. For the third model above, the initial candidate model consisted of all 34 variables in Table I, while the boundary model consisted of only the intercept. Variables were sequentially eliminated until there were either no variables remaining in the model or until there was no further reduction in the AIC.

2.3.2. Regression trees. Binary recursive partitioning methods were used to construct regression trees to predict 30-day AMI mortality [49, 50]. The R implementation of regression trees, like that of Breiman's *et al.*'s CART [49], only allows for binary partitions (or splits). In addition, the R implementation only allows for splits on individual variables and does not allow for splits on linear combinations of predictor variables. At each node, the split that maximizes the reduction in deviance is chosen.

Advantages to tree-based methods are that they do not require that one parametrically specify the nature of the relationship between the predictor variables and the outcome. Additionally, assumptions of linearity that are frequently made in linear and generalized linear models are not required for tree-based regression methods. Furthermore, tree-based methods are adept at identifying important interactions between predictor variables.

In the current study, an initial tree was grown using all 34 candidate predictor variables described in Table I. Once the initial regression tree had been grown, the tree was pruned. Ten-fold cross validation was used on the derivation data set to determine the optimal number of leaves on the tree [50]. The desired tree size was chosen to minimize the deviance when 10-fold cross validation was used on the derivation data set. Predictions were obtained on the validation data set using the pruned tree. The regression tree models were fit using the tree function in the TREE package for R. Pruning of the regression trees was done using the prune.tree function.

The following code was used in R:

```
tree.derive <- tree(mortality ~ x.1 + x.2 +···+ x.34, data=df.derive)        (1)
```

This grows a regression tree using binary recursive partitioning to predict mortality using the 34 variables listed in Table I. Node heterogeneity was measured using deviance. Ten-fold cross-validation was then done on the derivation data set as follows:

```
cv.tree.derive <- cv.tree(tree.derive,rand=1:10,K=10,FUN=prune.tree)        (2)
```

An optimal tree size was then selected to minimize the deviance. If two different tree sizes resulted in the same minimum deviance, then the smaller of the two tree sizes was chosen.

DOI: 10.1002/sim

This optimal tree size was denoted by best.size. The initial regression tree was pruned using the prune.tree function

```
prune.tree.derive <- prune.tree(tree.derive,best=best.size)        (3)
```

This final regression tree fit to the derivation sample was then used to obtain predictions for subjects in the validation sample.

2.3.3. Generalized additive models. A generalized additive model (GAM) is an additive regression model of the form

$$\eta(x) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) \tag{4}$$

where the f_i are functions of the predictors [51, 52]. The functions f_i can be either non-parametric scatterplot smoothers or regression splines. An advantage to the use of GAMs is that one can relax the linearity assumption between the predictor variable and the outcome. Furthermore, one does not have to specify the nature of the relationship, but can allow the data to dictate the nature of the relationship.

In the current study, we considered three separate GAMs for predicting 30-day AMI mortality. First, we considered the parsimonious regression model described above. Family history of CAD and presence of cardiogenic shock at admission were treated as binary predictor variables. Age, systolic blood pressure at admission, respiratory rate at admission, glucose, white blood count, and creatinine were modelled using smoothing splines, each with 5 degrees of freedom. A second GAM was fit that consisted of the above GAM, along with all two-way interactions. The third GAM contained all 34 variables listed in Table I. The 22 dichotomous variables were entered as binary predictor variables, while the 12 continuous variables were modelled using smoothing splines, each with 5 degrees of freedom. In the current study, smoothing terms were modelled using regression splines with 5 degrees of freedom for all three GAMs considered.

The GAMs were fit using the gam function in the GAM package in R. The first generalized additive was fit using the following code in R

```
gam(mortality ~ s(age,5) + s(sys.bp,5) + s(glucose,5) + s(wbc,5) + s(creatinine,5)
    + s(resp.rate,5) + cardiogenic.shock + fam.hx.cad, family=binomial, data=df.derive)
```

Similar code was used to fit the other two GAMs.

2.3.4. Multivariate adaptive regression spline models. Multivariate adaptive regression splines (MARS) is an adaptive regression procedure well suited to problems with a large number of predictor variables [53, 54]. MARS uses an expansion based on linear spline functions. For a given predictor variable X_j and a given value t taken by the predictor variable, one can define two linear spline functions: $(X_j - t)_+$ and $(t - X_j)_+$, where '$+$' refers to the positive part. A MARS model is constructed using a subset of all such possible linear spline functions. Furthermore, the products of such linear spline functions may also be considered. The reader is referred elsewhere for more details on variable selection and estimation [53, 54].

MARS is intended for continuous outcomes. Thus, 30-day mortality was initially modelled as a continuous outcome. The resultant design matrix obtained from the MARS analysis was then used as the design matrix in a logistic regression model with 30-day AMI mortality treated as a

DOI: 10.1002/sim

binary outcome. This method of using MARS to analyse dichotomous outcomes has been described by other authors [42]. We examined three separate MARS models. Each used the 34 variables described in Table I. The first model was an additive model that did not allow interactions between the predictor variables. The second model allowed for the inclusion of two-way interactions, while the third model allowed for the inclusion of all possible interactions, including 34-way interactions. MARS models are constructed using generalized cross-validation to determine the optimal number of terms in the model [54]. This use of generalized cross-validation helps protect against over-fitting the model in the derivation sample. Thus, while the third MARS model allowed for the potential inclusion of all possible interactions, the use of generalized cross-validation minimizes the likelihood that the final model will be over-fit to the derivation sample.

The MARS models were fit using the mars function in the MDA package. The following code was used to fit the first MARS model in R:

```
mars.derive <- mars(x.mat.pred, mortality, degree=1)
```

where x.mat.pred denotes a matrix of the predictor variables and mortality denotes the binary outcome variable. The MARS models that allowed for either two-way interactions or all possible interactions were fit similarly, by specifying degree=2 or degree=34, respectively.

2.4. *Comparison of predictive models*

Repeated split-sample validation was used to compare the predictive accuracy of each statistical method. The data were randomly divided into derivation and validation components. Two-thirds of the data were used for model derivation and the remaining one-third was used for model validation [55]. Each derivation sample consisted of 6323 subjects, while each validation sample consisted of 3161 subjects. This process was repeated 1000 times.

Each model was fit on the derivation sample. Predictions were then obtained for each subject in the validation sample using the model derived on the derivation sample. The predictive accuracy of each model was summarized by the area under the ROC curve [43]. This is equivalent to the *c*-statistic [43]. It has been suggested that the predictive ability of models be quantified using the ROC curve area [43]. The model area under the ROC curve was obtained for both the derivation and validation samples.

Harrell has suggested that the predictive ability of models can also be quantified using the generalized R_N^2 index of Nagelkerke [56] and Cragg and Uhler [57] and by Brier's score [43]. Brier's score is defined as

$$B = \frac{1}{n}\sum_{i=1}^{n} (\hat{P}_i - Y_i)^2 \tag{5}$$

where $\hat{P}_i$ is the predicted probability and $\hat{Y}_i$ is the observed response for the ith subject. While these indices are less commonly used than the area under the ROC curve, we include them in this study for comparative purposes. Accordingly, we computed the generalized R_N^2 index and Brier's score in each of the validation samples. The area under the ROC curve, the generalized R_N^2 index, and Brier's score were computed using the val.prob function from the design package for R. A limitation to using the area under the ROC curve is that it does not take into account the magnitude of the disagreement between the observed and predicted responses. The area under the ROC curve is equivalent to the proportion of all pairs consisting of one subject who experienced the outcome and one subject who does not experience the outcome in which the subject who experienced the

DOI: 10.1002/sim

outcome had a higher predicted probability of experiencing the outcome than does the subject who did not experience the outcome. As such, the magnitude of the disagreement is not taken into account. Indices such as Brier's score take into account the magnitude of the difference between observed and predicted responses.

Finally, model calibration in both the derivation and validation samples was assessed using the Hosmer–Lemeshow goodness-of-fit test [58]. This was done for all the modelling methods except for regression trees. The Hosmer–Lemeshow goodness-of-fit test divides the subjects into deciles of risk according the deciles of the predicted probability of 30-day mortality. However, all regression trees had fewer than 10 terminal nodes, thus one could not partition the predicted probabilities of 30-day mortality into deciles.

The above methods were repeated 1000 times: the initial data were randomly divided into derivation and validation components 1000 times. Each predictive model was fit using the derivation data set and predictions were then obtained on the validation data set. Results were then summarized over the 1000 validation data sets. By using 1000 different derivation/validation samples, we were able to assess the robustness of our results under different derivation and validation samples.

3. RESULTS

3.1. Predictive performance

The mean area under the ROC curve for each model in both the 1000 derivation and validation samples is reported in Table II. In the validation sample, the mean ROC curve area for the regression tree model was 0.762, while the mean ROC curve area for the simple logistic regression model

Table II. Model calibration and discrimination in the 1000 repeated split samples.

Model	ROC area: derivation sample	ROC area: validation sample	Hosmer–Lemeshow GOF: derivation sample	Hosmer–Lemeshow GOF: validation sample	R_N^2: validation sample	Brier's score: validation sample
Regression tree	0.779	0.762			0.198	0.087
Logistic regression (eight main effects)	0.846	0.845	0.2271	0.2363	0.319	0.078
Logistic regression (two-way interactions)	0.849	0.844	0.2255	0.2109	0.313	0.078
Logistic regression (backwards elimination from full model)	0.853	0.846	0.2243	0.2137	0.321	0.078
GAM (eight main effects)	0.857	0.850	0.3642	0.2493	0.333	0.076
GAM (two-way interactions)	0.861	0.849	0.5526	0.1984	0.328	0.077
GAM (full model)	0.869	0.851	0.2263	0.1316	0.332	0.077
MARS (additive)	0.858	0.848	0.0820	0.1139	0.326	0.077
MARS (two-way interactions)	0.867	0.837	0.0947	0.0167	0.275	0.080
MARS (all interactions)	0.868	0.831	0.0748	0.0051	0.244	0.082

Note: Results are averaged over the 1000 derivation and validation samples.

DOI: 10.1002/sim

was 0.845. The difference in ROC curve areas for the regression tree method and the simple logistic regression model ranged from a low of 0.041 to a high of 0.165 across the 1000 derivation samples (mean difference: 0.083). The mean ROC curve areas for the other modelling methods in the validation samples ranged from a low of 0.831 (MARS model that allowed for the inclusion of all possible interactions) to a high of 0.851 (GAM consisting of all main effects). The two MARS models that incorporated interactions had lower predictive accuracy than the three logistic regression models, the three GAMs, and the additive MARS model. This may be indicative that the MARS models that allowed for the inclusion of interactions were over-fit on the derivation samples. Both paired t-tests and Wilcoxon matched-pairs signed-ranks tests were used to compare the ROC curve of the regression tree models with that of each of the other predictive models. Each competing model had a significantly higher ROC curve area than the regression tree models ($P < 0.0001$ for both tests and for each pair-wise comparison).

The mean ROC curve area for the regression tree model decreased from 0.779 in the derivation samples to 0.762 in the validation samples—a decrease of 0.017. The decline in the mean ROC curve area between the derivation and validation samples was negligible for the simple logistic regression model (0.846 to 0.845). Similarly, the drop in ROC curve area from the derivation sample to the validation sample was negligible for the logistic regression model that included two-way interactions (0.849 to 0.844), the logistic regression model obtained using backward elimination (0.853 to 0.846), for the simple GAM (0.857 to 0.850). For the remaining models, the decline in ROC curve areas from the derivation samples to the validation samples ranged from 0.010 (additive MARS model) to 0.037 (MARS models allowing for all interactions). The greater decline in ROC curve area for the MARS models compared to the simpler logistic regression and GAMs may be indicative of a tendency of the more complex MARS models to be over-fit on the derivation samples. For most models, the decrease in ROC curve area from the derivation sample to the validation sample was relatively modest. This may be attributable to the relatively large number of subjects in the derivation samples (6323 subjects). It is possible that each method would display greater over-optimism in the derivation samples if the derivation samples were smaller. The MARS models that allowed for the inclusion of either two-way interactions or for the inclusion of all possible interactions demonstrated modest over-optimism in the derivation samples. However, there is no evidence that these two model results were substantially over-fit in the derivation samples. This is likely attributable to the fact that MARS employs generalized cross-validation in selecting terms for inclusion in the regression model.

The distribution of the area under the ROC curve in the 1000 validation data sets for each modelling approach is described in Figure 1. In order to improve the resolution of the figure, we truncated the display of the figure to display only the portion of the density plots that lay above 0.69 (the MARS model with two-way interactions had one observed ROC curve value less than this threshold, while the MARS model that allowed for all possible interactions had three ROC curve values less than this threshold). Several observations are apparent. First, the distribution of ROC curve areas for the regression tree models was shifted downwards compared to that of the other modelling approaches. By using 1000 derivation/validation samples, we demonstrated that the regression tree models had consistently poorer performance than the other methods. Second, the distributions of ROC curve areas for the simple logistic regression model, the logistic regression model that included two-way interactions, and the logistic regression model obtained from the full model using backwards elimination were almost identical. Third, the same phenomenon was observed for the simple GAM and for the GAM that included interactions. Fourth, while there was

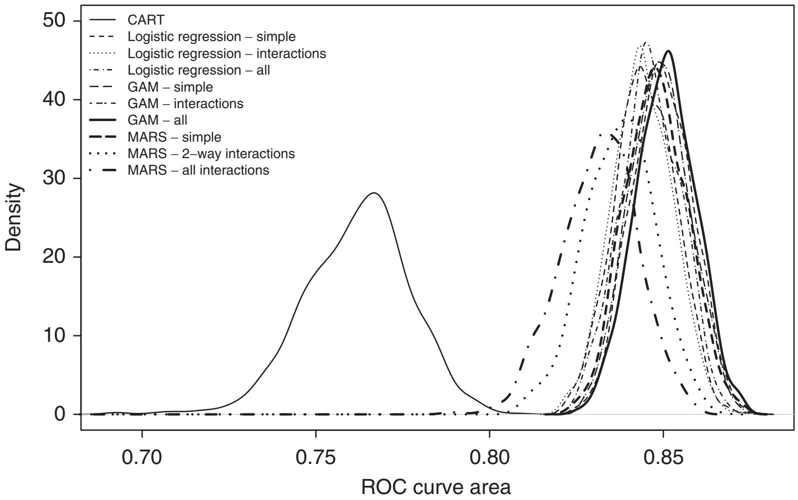

Figure 1. Distribution of ROC curve areas in 1000 validation samples.

some variability between the non-regression tree methods, the greatest difference was between the regression tree models and all the other methods. Fifth, the two MARS models that allowed for the incorporation of interactions exhibited greater variability in ROC curve areas in the validation samples than did the use of logistic regression, GAMs, or the additive MARS model. Sixth, our use of repeated use of split-sample validation allowed us to describe the distribution of the ROC curve area in the validation samples. While the mean ROC curve area for the regression tree method was 0.762, the observed ROC curves in the derivation samples ranged from a low of 0.692 to a high of 0.808. Earlier studies that did not use repeated split-sample validation were unable to examine the degree to which their results depended, in part, on the particular manner in which the sample was split into derivation and validation samples.

The generalized R_N^2 index is reported in Table II for each of the modelling strategies. The index ranged from a low of 0.198 for the regression tree model to a high of 0.333 for the simple GAM. The generalized R_N^2 indices for the three logistic regression models, the three GAMs and the additive MARS model were comparable (0.313 to 0.333). Of note is the fact that the regression tree model explained a substantially smaller proportion of the observed variation than did the other regression methods. The Pearson and Spearman correlation coefficients for the correlation between the generalized R_N^2 indices and the area under the ROC curves in the validation samples were 0.90 and 0.99, respectively, indicating good concordance between these two measures of predictive ability. The Brier's score is reported in Table II for each of the modelling strategies. The Brier's scores ranged from a high of 0.087 for the regression tree model to a low of 0.076 for the simple GAM. The Brier's scores for the three logistic regression models, the three GAMs and the additive MARS model were comparable (0.076 to 0.078). The distribution of the generalized R_N^2 index and Brier's scores in the 1000 validation data sets for each modelling approach are described in Figures 2 and 3, respectively. Similar observations can be made concerning the distribution of the generalized R_N^2 index and Brier's scores in the validation samples as was made above for the distribution of the ROC curve areas for the different models in the validation samples.

DOI: 10.1002/sim

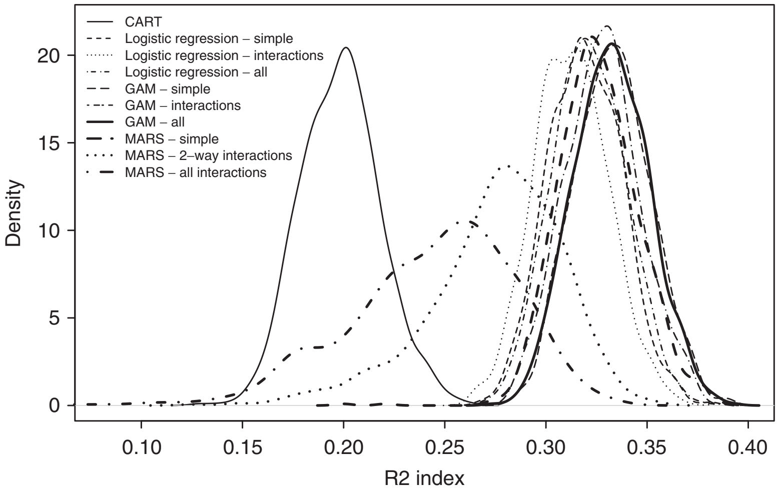

Figure 2. Distribution of R^2 index in 1000 validation samples.

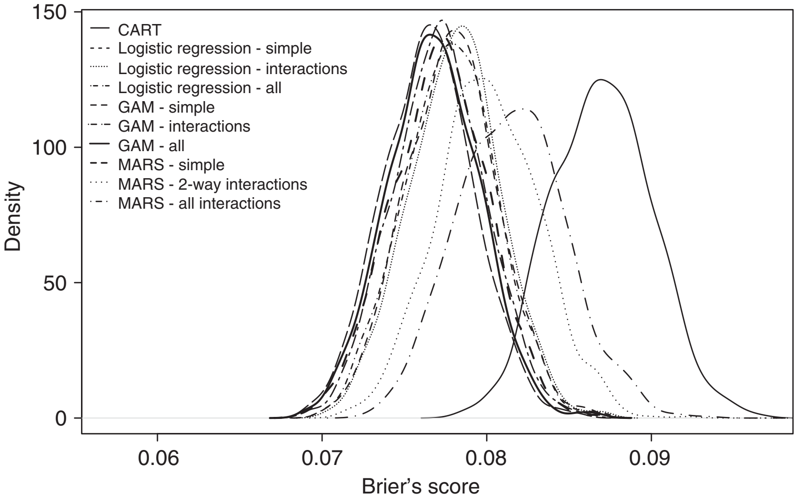

Figure 3. Distribution of Brier's scores in 1000 validation samples.

3.2. Model calibration

The results for the Hosmer–Lemeshow goodness-of-fit test are reported in Table II. On average, the logistic regression and GAMs fit the derivation and validation sample samples well ($P>0.22$ for derivation sample; $P>0.13$ for validation sample). The additive MARS model showed some evidence of lack of fit ($P=0.0820$ and 0.1139 in the derivation and validation samples, respectively). Both the MARS model that allowed for the inclusion of two-way interactions and the MARS model that allowed for the inclusion of all possible interactions showed evidence of

DOI: 10.1002/sim

poorly fitting the data ($P = 0.0947$ and 0.0748 in the derivation samples, respectively; $P = 0.0167$ and 0.0051 in the validation samples, respectively). The Hosmer–Lemeshow goodness-of-fit test was not reported for the regression tree models since in many instances it was not possible to divide subjects into deciles of risk. This was due to the regression trees having fewer than 10 terminal nodes, and thus there was insufficient variability in predicted probabilities to compute deciles of risk.

3.3. Miscellaneous results

The mean number of terminal nodes for the regression trees across the 1000 derivation samples was 6.4. The number of terminal nodes ranged from a low of 4 to a high of 9. The first and third quartiles were 6 and 7, respectively. The number of variables used in constructing the regression trees ranged from a low of 3 to a high of 6, with a mean of 4.4. The first and third quartiles were 4 and 5, respectively. The regression tree obtained using one of the derivation samples is illustrated in Figure 4. This particular regression tree had 7 terminal nodes. Five variables were

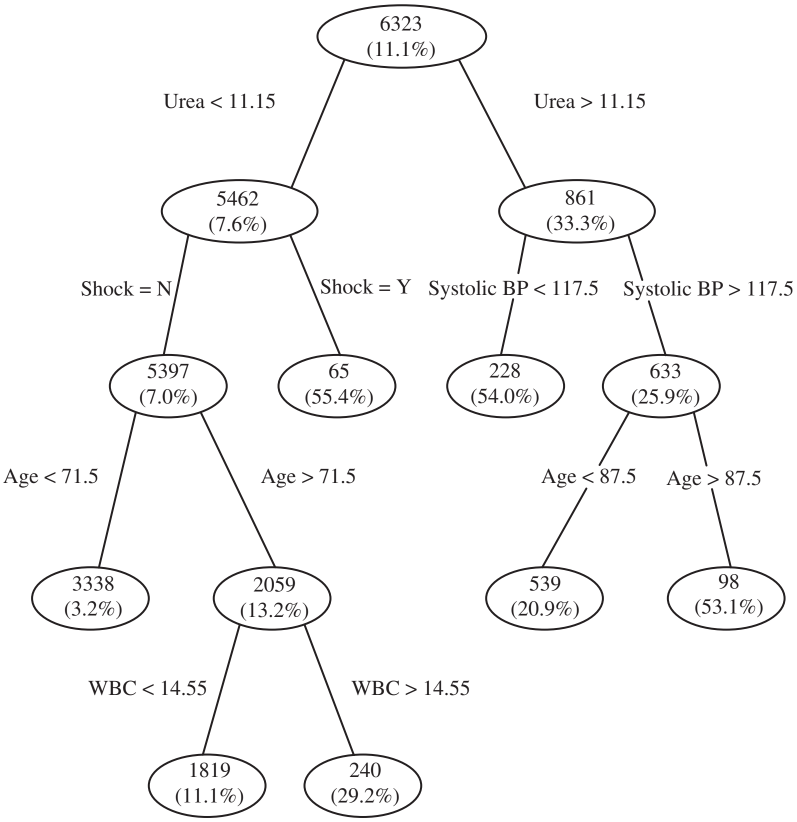

Figure 4. Regression tree for 30-day mortality. Each node contains the number of subjects in that node and the 30-day mortality rate of those subjects [*N* (30-day mortality rate)].

DOI: 10.1002/sim

used in creating the tree: urea, age, white blood count, cardiogenic shock at admission, and systolic blood pressure.

The mean number of interactions identified in the 1000 derivation samples using the logistic regression model that allowed for the incorporation of interactions was 7.8. The number of interactions ranged from a low of 2 to a high of 15. The first and third quartiles were 6 and 9, respectively. The mean number of main effects identified in the 1000 derivation samples using the logistic regression model that used backwards elimination on the full model was 17.3. The number of main effects varied from a low of 12 to a high of 23. The first and third quartiles were 16 and 18, respectively.

The mean number of parameters in the MARS models that included only main effects was 20.1 (range: 14–29). The first and third quartiles were 18 and 22, respectively. The mean number of terms in the MARS models that allowed the inclusion of two-way interactions was 42.0 (range: 23–53). The first and third quartiles were 40 and 44, respectively. The mean number of terms in the MARS models that allowed for the inclusion of all possible interactions was 43.2 (range: 34–52). The first and third quartiles were 41 and 45, respectively.

Figure 5 describes the relationship between the log-odds of 30-day mortality and age, systolic blood pressure, glucose level, white blood count, creatinine level, and respiratory rate (the six continuous variables used in the first GAM) derived from the first GAM in each of the first 20 derivation samples (this method of depicting variability in models across samples has been used in earlier studies [59]). For each value of a given predictor variable, we determined the predicted log-odds of 30-day mortality, holding the other continuous variables fixed at the sample average and the dichotomous predictor variables set to absent. At the base of each graph is a jittered rug-plot, depicting the distribution of each predictor in the overall sample. We then fit the above

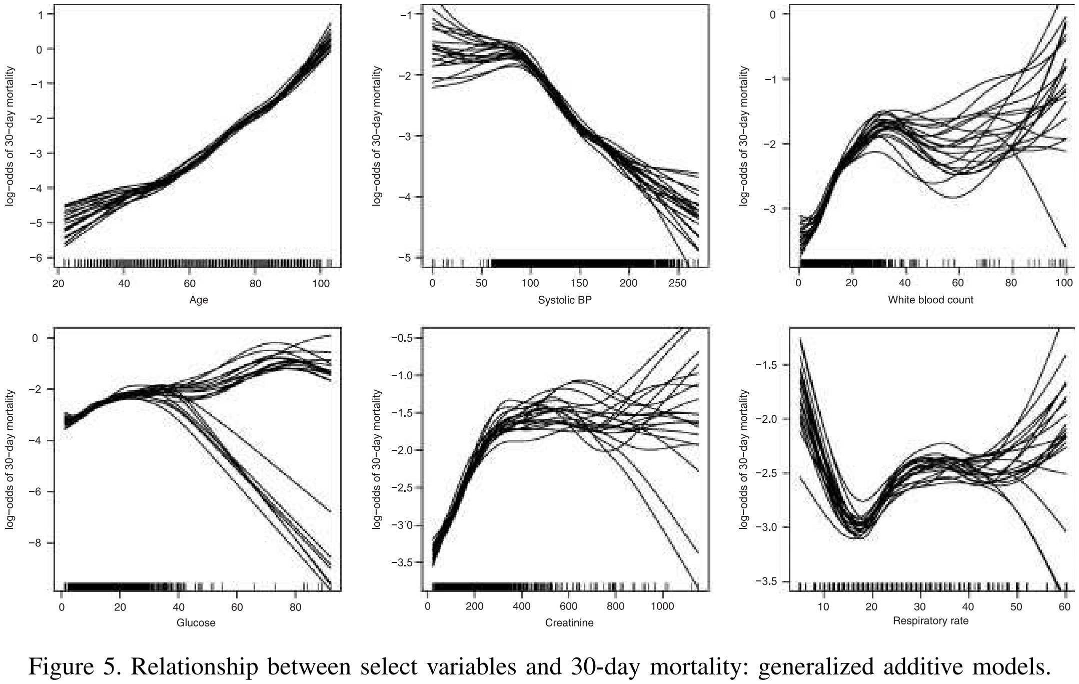

Figure 5. Relationship between select variables and 30-day mortality: generalized additive models.

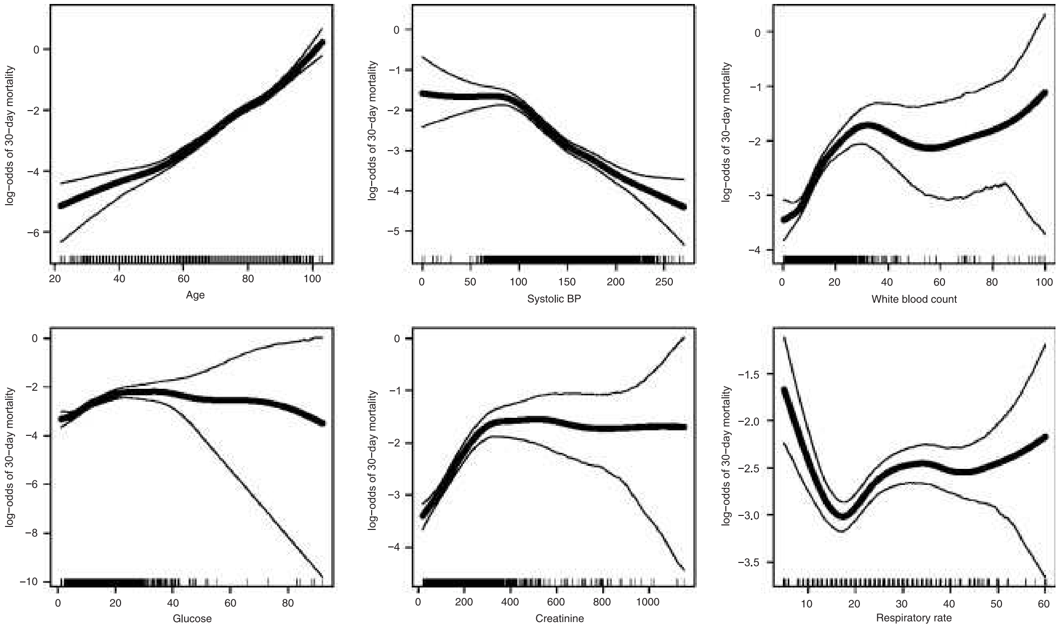

Figure 6. Mean relationship between select variables and 30-day mortality: generalized additive models.

GAM to each of the 1000 derivation samples. Allowing each of the six variables to incrementally increase across the observed range for that variable in the sample, we then computed the mean log-odds of 30-day mortality and the 2.5th and 97.5th percentiles of the log-odds of 30-day mortality across the 1000 models fit to the derivation samples. The relationship between age, systolic blood pressure, glucose level, white blood count, creatinine level, respiratory rate and the mean log-odds of 30-day mortality is described in Figure 6, along with the empirical 2.5th and 97.5th percentiles of the predicted log-odds of 30-day mortality across the 1000 derivation samples. One observes that the relationship between age and the log-odds of 30-day mortality is approximately linear over the entire range of observed ages. For the remaining five continuous variables, nonlinear relationships were evident. However, with the exception of respiratory rate, the relationships were predominately linear over a range of the distribution in which a majority of subjects lay (see Table I for 25th and 75th percentiles of each distribution). For systolic blood pressure, there was an initially flat relationship with the mean log-odds of 30-day mortality. However, following this initial plateau, a negative linear relationship with 30-day mortality was observed. For creatinine, a linear relationship with 30-day mortality was initially observed. However, this relationship attained a plateau, and it appears that further increases in creatinine do not have an impact on 30-day AMI mortality. For respiratory rate, a complex nonlinear relationship with 30-day mortality was observed. In particular, this nonlinear relationship was evident over the range of the distribution in which the large majority of subjects lay. Additionally, as depicted in Figure 5, this relationship was observed in each of the first 20 derivation samples. Figures 5 and 6 are complementary. Figure 5 demonstrates how individual regression relationships vary across derivation samples, while Figure 6 demonstrates the average regression relationship across the 1000 derivation samples.

DOI: 10.1002/sim

This average relationship is less likely to be influenced by influential observations that may be included or omitted in certain derivation samples. In examining the rug-plot at the base of each figure, one observes that extreme data values can be observed for some of the continuous variables. For instance, there are subjects with systolic blood pressure lower than 20. Patients presenting in cardiogenic shock can have very low or non-existent systolic blood pressure. Since 1.6 per cent of the sample presented cardiogenic shock (Table I), it is to be expected that there will be a minority of patients with very low or absent systolic blood pressure. The chart abstraction instrument for abstracting the data from AMI patient's medical records incorporated built-in range checks. The upper and lower endpoints of these ranges were obtained through consultation with experts in cardiovascular and internal medicine. Thus, while some of the observed values are large, they are within the plausible range for subjects who have experienced an AMI.

Figure 7 describes the relationship between the log-odds of 30-day mortality and age, systolic blood pressure, glucose level, white blood count, creatinine level, and respiratory rate as derived from the additive MARS model with no interactions fit to each of the first derivation 20 samples. In most of the 20 samples, age was linearly related to the log-odds of 30-day mortality. In one of the 20 samples, there was no independent effect of age on mortality from ages 20 to 60. However, from age 60 onwards, the risk of 30-day mortality increased linearly. In most of the 20 samples, there was a linear relationship between systolic blood pressure and the log-odds of 30-day mortality. Two variables (glucose level and creatinine level) displayed a linear relationship with mortality, which then reached a plateau. The relationship between white blood count and

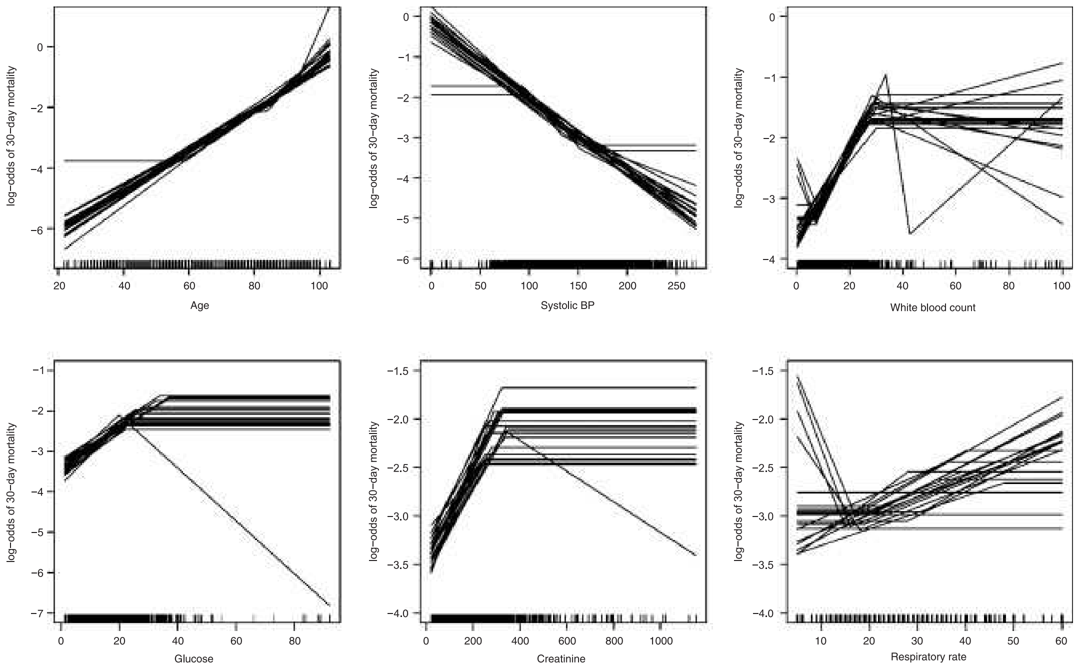

Figure 7. Relationship between select variables and 30-day mortality: multivariate adaptive regression splines.

DOI: 10.1002/sim

mortality was linear over the range of the distribution in which the large majority of subjects lay. The relationship between respiratory rate and the log-odds of 30-day mortality was more inconsistent across the 20 derivation samples. In some samples the relationship was linear, where as in other samples a V-shaped relationship was evident. We then fit the additive multivariate adaptive regression spline model to each of the 1000 derivation samples. For each value of the given predictor variables, we determined the predicted log-odds of 30-day mortality, holding the other continuous variables fixed at the sample average, and the dichotomous independent variables set to absent. Allowing each of the six variables to incrementally increase across the observed range for that variable in the sample, we then computed the mean log-odds of 30-day mortality and the 2.5th and 97.5th percentiles of the log-odds of 30-day mortality across the 1000 models fit to the derivation samples. The relationship between age, systolic blood pressure, glucose level, white blood count, creatinine level, respiratory rate, and the mean log-odds of 30-day mortality is described in Figure 8, along with the associated empirical 95 per cent prediction intervals. As in Figure 6, one observes that age and systolic blood pressure are linearly related to the log-odds of 30-day mortality across the range of age and systolic blood pressure observed in the study sample. White blood count, glucose, and creatinine display a linear relationship with the log-odds of 30-day mortality that attains a plateau. However, for each of these three variables, a linear relationship is evident over a range in which a majority of the subjects lay. The relationship between respiratory rate and the log-odds of 30-day mortality is V-shaped. The relationships between the log-odds of 30-day mortality and age, systolic blood pressure, white blood count, and glucose described

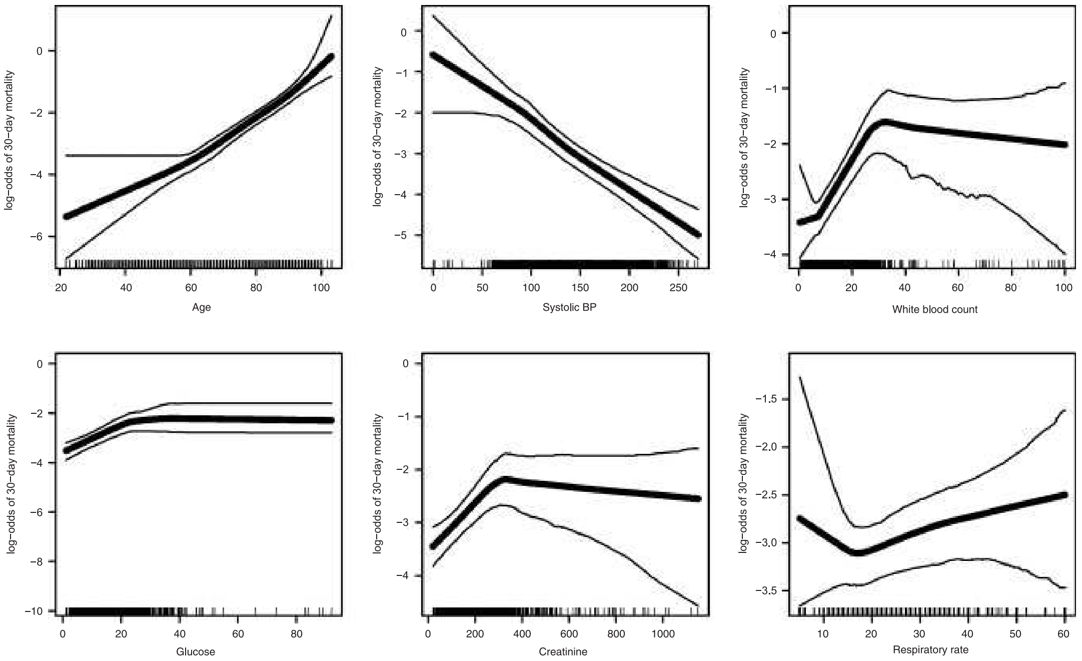

Figure 8. Mean relationship between select variables and 30-day mortality: multivariate adaptive regression splines models.

DOI: 10.1002/sim

in Figure 8 are qualitatively similar to those depicted in Figure 6 over a range of the variable in which the majority of the subjects lay.

4. DISCUSSION

There is an increasing interest in using regression trees to identify subgroups of patients at increased risk for adverse events and outcomes. In the current study, we have demonstrated, using a large sample of patients hospitalized with AMI, that regression tree methods did not predict 30-day mortality as accurately as did conventional logistic regression. Furthermore, we demonstrated that the predictive performance of conventional logistic regression was comparable to that of modern flexible regression methods such as GAMs and MARS models. The mean ROC curve area for conventional logistic regression with no interactions was 0.845 in the validation sample, while the corresponding value for the regression tree model was 0.762. The ROC curve areas were at most slightly higher for more flexible, data-driven regression methods than for the simple logistic regression model. The highest observed ROC curve area was for a GAM that consisted of all 34 main effects (ROC curve area 0.851). In addition to comparing the predictive accuracy of regression models from different families of models, we also compared the predictive accuracy of models with differing complexity from the same family of models. We found that more complex models from the same family had at best negligibly higher predictive accuracy (logistic regression and GAMs) or lower predictive accuracy (MARS). Similar results were observed when the generalized R_N^2 index and Brier's score were used to quantify the predictive accuracy of different regression models.

Several explanations exist for our findings. First, analyses conducted using GAMs indicated that the relationship between the log-odds of 30-day mortality and age, systolic blood pressure, glucose, white blood count, and creatinine was approximately linear over the range of the majority of the distribution for each of these variables. This finding was confirmed by the additive MARS model, which indicated that the relationship between several predictors and the outcome was linear over the majority range of each predictor variable. Thus, the conventional logistic model was able to exploit the strong underlying linear relationships in the data. However, the regression tree model relies on partitioning the sample using binary decision rules, and thus loses the ability to incorporate underlying linear but not piecewise constant relationships. Second, one of the described advantages of regression tree methods is their ability to identify interactions among the predictor variables. We fit two separate logistic regression models: one with only main effects and one that incorporated two-way interactions. The predictive performance of the two models in the validation samples was almost identical. Additionally, we fit two separate GAMs: one with only main effects, and one that incorporated two-way interactions. As before, the predictive performance of the two models in the validation samples was almost identical. Similarly, the predictive performance of the MARS model that only included main effects was greater than that of either MARS model that allowed for possible interactions. These sets of analyses indicate that, while important interactions may exist, their inclusion does not improve the prediction of AMI mortality. Regression trees have difficulty in capturing additive relationships [54]. This may have contributed to the poor performance of these methods in our sample. The GAM consisting of only eight variables demonstrated that the existence of a complex nonlinear relationship between respiratory rate the log-odds of 30-day mortality. The nonlinear nature of this relationship may explain why the GAMs had marginally higher predictive ability than the conventional logistic regression models.

DOI: 10.1002/sim

We offer the following two suggestions to researchers assessing the predictive accuracy of specific regression models or comparing the predictive accuracy of different regression models. First, the predictive accuracy of regression models should be assessed using a summary measure such as the area under the ROC curve (equivalent to the *c*-statistic) in an independent validation sample. While most of the articles summarized in the Introduction used independent derivation and validation samples, many authors relied upon arbitrarily dichotomizing the predicted probability of an event, and then computing sensitivity and specificity. As stated in the Introduction, this approach has been criticized for a variety of reasons [43, 44]. Second, repeated split-sample validation should be employed. For each regression model, we observed meaningful variability in the ROC curve area across the 1000 validation samples. To the best of our knowledge, only one study has compared CART with logistic regression using repeated split sample validation to examine the robustness of the findings to the particular splitting of the sample in derivation and validation samples [38]. Repeated split-sample validation allows one to examine variation in the predictive accuracy of a given regression model. Furthermore, the use of repeated split-sample derivation samples allows one to develop a more robust characterization of the nature of the relationship between specific predictor variables and the outcome. Using repeated split-sample validation, we were able to determine the mean regression relation across the 1000 derivation samples. This allows one to determine a mean relationship that is less likely to be influenced by influential observations that may be included in only a few derivation samples.

In conclusion, we have demonstrated that logistic regression had superior predictive ability compared to regression trees for predicting AMI mortality. Furthermore, modern data-driven methods such as GAMs and MARS models had predictive ability that was at best only modestly greater than that of conventional logistic regression. The methods outlined in this study will allow researchers to comprehensively compare the predictive accuracy of different regression methods.

ACKNOWLEDGEMENTS

The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported by operating grant from the Canadian Institutes of Health Research (CIHR). Dr Austin is supported in part by a New Investigator award from the CIHR. The data used in this study were obtained from the EFFECT study. The EFFECT study was funded by a Canadian Institutes of Health Research (CIHR) Team Grant in Cardiovascular Outcomes Research. The author would like to thank two anonymous reviewers whose comments substantially improved the manuscript.

REFERENCES

1. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *Journal of the American Medical Association* 2003; **290**:2581–2587.
2. Tu JV, Jaglal SB, Naylor CD *et al*. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. *Circulation* 1995; **91**:677–684.
3. Lee KL, Woodlief LH, Topol EJ *et al*. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. *Circulation* 1995; **91**:1659–1668.
4. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: the Framingham study risk score functions. *Statistics in Medicine* 2004; **23**:1631–1660.

DOI: 10.1002/sim

5. Iezzoni LI. *Risk Adjustment for Measuring Healthcare Outcomes* (2nd edn). Health Administration Press: Chicago, IL, 1997.
6. Geelhoed M, Boerrigter AO, Camfield P, Geerts AT, Arts W, Smith B, Camfield C. The accuracy of outcome prediction models for childhood-onset epilepsy. *Epilepsia* 2005; **46**:1526–1532.
7. Nishida N, Tanaka M, Hayashi N, Nagata H, Takeshita T, Nakayama K, Morimoto K, Shizukuishi S. Determination of smoking and obesity as periodontitis risks using the classification and regression tree method. *Journal of Periodontology* 2005; **76**:923–928.
8. Dessein PH, Joffe BI, Veller MG, Stevens BA, Tobias M, Reddi K, Stanwix AE. Traditional and nontraditional cardiovascular risk factors are associated with atherosclerosis in rheumatoid arthritis. *Journal of Rheumatology* 2005; **32**:435–442.
9. Kuchibhatla M, Fillenbaum GG. Alternative statistical approaches to identifying dementia in a community-dwelling sample. *Aging and Mental Health* 2003; **7**:383–389.
10. Avila PC, Segal MR, Wong HH, Boushey HA, Fahy JV. Predictors of late asthmatic response. Logistic regression and classification tree analyses. *American Journal of Respiratory and Critical Care Medicine* 2000; **161**: 2092–2095.
11. Bhavnani SM, Drake JA, Forrest A, Deinhart JA, Jones RN, Biedenbach DJ, Ballow CH. A nationwise, multicenter, case–control study comparing risk factors, treatment, and outcome for vancomycin-resistant and -susceptible enterococcal bacteremia. *Diagnostic Microbiology and Infectious Disease* 2000; **36**:145–158.
12. Wietlisbach V, Vader JP, Porchet F, Costanza MC, Burnand B. Statistical approaches in the development of clinical practice guidelines from expert panels: the case of laminectomy in sciatica patients. *Medical Care* 1999; **37**:785–797.
13. Roehrborn CG, Malice M, Cook TJ, Girman CJ. Clinical predictors of spontaneous acute urinary retention in men with LUTS and clinical BPH: a comprehensive analysis of the pooled placebo groups of several large clinical trials. *Urology* 2001; **58**:210–216.
14. Knuiman MW, Vu HT, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *Journal of Cardiovascular Risk* 1997; **4**:127–134.
15. Litvan I, Campbell G, Mangone CA, Verny M, McKee A, Chaudhuri KR, Jellinger K, Pearce RK, D'Olhaberriague L. Which clinical features differentiate progressive supranuclear palsy (Steele–Richardson–Olszewski syndrome) from related disorders? A clinicopathological study. *Brain* 1997; **120**:65–74.
16. Pilote L, Miller DP, Califf RM, Rao JS, Weaver WD, Topol EJ. Determinants of the use of coronary angiography and revascularization after thrombolysis for acute myocardial infarction. *New England Journal of Medicine* 1996; **335**:1198–1205.
17. Bhattacharyya S, Siegel ER, Petersen GM, Chari ST, Suva LJ, Haun RS. Diagnosis of pancreatic cancer using serum proteomic profiling. *Neoplasia* 2004; **6**:674–686.
18. Taylor WJ, Marchesoni A, Arreghini M, Sokoll K, Helliwell PS. A comparison of the performance characteristics of classification criteria for the diagnosis of psoriatic arthritis. *Seminars in Arthritis and Rheumatism* 2004; **34**:575–584.
19. Seligman DA, Pullinger AG. Improved interaction models of temporomandibular joint anatomic relationships in asymptomatic subjects and patients with disc displacement with or without reduction. *Journal of Orofacial Pain* 2004; **18**:192–202.
20. Stalans LJ, Yarnold PR, Seng M, Olson DE, Repp M. Identifying three types of violent offenders and predicting violent recidivism while on probation: a classification tree analysis. *Law and Human Behavior* 2004; **28**:253–271.
21. Arena VC, Sussman NB, Mazumdar S, Yu S, Macina OT. The utility of structure-activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision three models. *Sar and Qsar in Environmental Research* 2004; **15**:1–18.
22. Schwarzer G, Nagata T, Mattern D, Schmelzeisen R, Schumacher M. Comparison of fuzzy inference, logistic regression, and classification trees (CART). Prediction of cervical lymph node metastasis in carcinoma of the tongue. *Methods of Information in Medicine* 2003; **42**:572–577.
23. Thwaites GE, Chau TT, Stepniewska K, Phu NH, Chuong LV, Sinh DX, White NJ, Parry CM, Farrar JJ. Diagnosis of adult tuberculous meningitis by use of clinical and laboratory features. *Lancet* 2002; **360**:1287–1892.
24. Pullinger AG, Seligman DA, John MT, Harkins S. Multifactorial modeling of temporomandibular anatomic and orthopedic relationships in normal versus undifferentiated disk displacement joints. *Journal of Prosthetic Dentistry* 2002; **87**:289–297.
25. Pullinger AG, Seligman DA. Multifactorial analysis of differences in temporomandibular joint hard tissue anatomic relationships between disk displacement with and without reduction in women. *Journal of Prosthetic Dentistry* 2001; **86**:407–419.

DOI: 10.1002/sim

26. Seligman DA, Pullinger AG. Analysis of occlusal variables, dental attrition, and age for distinguishing healthy controls from female patients with intracapsular temporomandibular disorders. *Journal of Prosthetic Dentistry* 2000; **83**:76–82.
27. Woolas RP, Conaway MR, Xu F, Jacobs IJ, Yu Y, Daly L, Davies AP, O'Briant K, Berchuck A, Soper JT *et al*. Combinations of multiple serum markers are superior to individual assays for discriminating malignant from benign pelvic masses. *Gynecologic Oncology* 1995; **59**:111–1116.
28. Li D, German D, Lulla S, Thomas RG, Wilson SR. Prospective study of hospitalization for asthma. A preliminary risk factor model. *American Journal of Respiratory and Critical Care Medicine* 1995; **151**:647–655.
29. Hasford J, Ansari H, Lehmann K. CART and logistic regression analyses of risk factors for first dose hypotension by an ACE-inhibitor. *Therapie* 1993; **48**:479–482.
30. Stewart PW, Stamm JW. Classification tree prediction models for dental caries from clinical, microbiological, and interview data. *Journal of Dental Research* 1991; **70**:1239–1251.
31. Sauerbrei W, Madjar H, Prompeler HJ. Differentiation of benign and malignant breast tumors by logistic regression and a classification tree using Doppler flow signals. *Methods of Information in Medicine* 1998; **37**:226–234.
32. Rosenfeld B, Lewis C. Assessing violence risk in stalking cases: a regression tree approach. *Law and Human Behavior* 2005; **29**:343–357.
33. Garzotto M, Beer TM, Hudson RG, Peters L, Hsieh YC, Barrera E, Klein T, Mori M. Improved detection of prostate cancer using classification and regression tree analysis. *Journal of Clinical Oncology* 2005; **23**:4322–4329.
34. Gansky SA. Dental data mining: potential pitfalls and practical issues. *Advances in Dental Research* 2003; **17**:109–114.
35. El-Solh AA, Sikka P, Ramadan F. Outcome of older patients with severe pneumonia predicted by recursive partitioning. *Journal of the American Geriatrics Society* 2001; **49**:1614–1621.
36. Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. *Medinfo* 1998; **9**:493–497.
37. 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. *Journal of Investigative Medicine* 1995; **43**:468–476.
38. James KE, White RF, Kraemer HC. Repeated split sample validation to assess logistic regression and recursive partitioning: an application to the prediction of cognitive impairment. *Statistics in Medicine* 2005; **24**:3019–3035.
39. Long WJ, Griffith JL, Selker HP, D'Agostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research* 1993; **26**:74–97.
40. Nelson LM, Bloch DA, Longstreth Jr WT, Shi H. Recursive partitioning for the identification of disease risk subgroups: a case–control study of subarachnoid hemorrhage. *Journal of Clinical Epidemiology* 1998; **51**:199–209.
41. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine* 2003; **26**:172–181.
42. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. *Statistics in Medicine* 1998; **17**:2501–2508.
43. Harrell Jr FE. *Regression Modeling Strategies*. Springer: New York, 2001.
44. van Houwelingen JC, le Cessie S. Predictive value of statistical models. *Statistics in Medicine* 1990; **8**:1303–1325.
45. Tu JV, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, Newman AM. *Quality of Cardiac Care in Ontario*. Institute for Clinical Evaluative Sciences: Toronto, Ont., 2004.
46. Austin PC, Tu JV. Bootstrap methods for developing predictive models in cardiovascular research. *The American Statistician* 2004; **58**:131–137.
47. R Core Development Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing: Vienna, 2005.
48. Hastie TJ, Pregibon D. Generalized linear models. In *Statistical Models in S*, Chambers JM, Hastie TJ (eds). Chapman & Hall: New York, 1993.
49. Breiman L, Freidman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall/CRC Press: London, Boca Raton, FL, 1998.
50. Clark LA, Pregibon D. Tree-based methods. In *Statistical Models in S*, Chambers JM, Hastie TJ (eds). Chapman & Hall: New York, 1993.
51. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: London, 1990.
52. Hastie TJ. Generalized additive models. In *Statistical Models in S*, Chambers JM, Hastie TJ (eds). Chapman & Hall: New York, 1993.

DOI: 10.1002/sim

53. Friedman JH. Multivariate adaptive regression splines. *The Annals of Statistics* 1991; **19**:1–67.
54. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Data Mining*, *Inference*, *and Prediction*. Springer: New York, 2001.
55. Picard RR, Berk KN. Data splitting. *The American Statistician* 1990; **44**:140–147.
56. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika* 1991; **78**:691–692.
57. Cragg JG, Uhler R. The demand for automobiles. *Canadian Journal of Economics* 1970; **3**:386–406.
58. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 1989.
59. Austin PC, Escobar MD. The use of finite mixture models to estimate the distribution of the health utilities index in the presence of a ceiling effect. *Journal of Applied Statistics* 2003; **30**:909–923.

DOI: 10.1002/sim