19

Design of Scoring Models for Trustworthy Risk Prediction in Critical Patients

Paolo Barbini and Gabriele Cevenini
Department of Surgery and Bioengineering, University of Siena
Italy

1. Introduction

Prediction of an adverse health event (AHE) from objective data is of great importance in clinical practice. A health event is inherently dichotomous as it either happens or does not happen, and in the latter case, it is a favourable health event (FHE).

In many clinical applications, it is relevant not only to predict AHEs happening (diagnostic ability) but also to estimate in advance their individual risk of occurrence using ordered multinomial or quantitative scales (prognostic ability) such as probability. An estimated probability of a patient's outcome is usually preferred to a simpler binary decision rule. However, models cannot be designed by optimising their fit to true individual risk probabilities because the latter are not intrinsically known, nor can they be easily and accurately associated with an individual's data. Classification models are therefore usually trained on binary outcomes to provide an orderable or quantitative output, which can be dichotomised using a suitable cut-off value.

Model discrimination refers to accurate identification of actual outcomes. Model calibration, or goodness of fit, is related to the agreement between predicted probabilities and observed proportions and it is an important aspect to consider in evaluating the prognostic capacity of a risk model (Cook, 2008). Model calibration is independent of discrimination, since there are risk models with good discrimination but poor calibration. A well-calibrated model gives probability values that can be reliably associated with the true individual risk of outcomes.

Many models have recently been proposed for diagnostic purposes in a wide range of medical applications and they also provide reliable estimates of individual risk probabilities. Two different approaches have been used to predict patient risk. The first approach is based on estimation of risk probability by sophisticated mathematical and statistical methods, such as logistic regression, the Bayesian rule and artificial neural networks (Dreiseitl & Ohno-Machado, 2002; Fukunaga, 1990; Marshall et al., 1994). Despite their great accuracy, these models are unfortunately not widely used because they are hard to design and call for difficult calculations, often requiring dedicated software and computing knowledge that doctors do not welcome, besides being difficult to incorporate in clinical practice. The second approach creates scoring systems, in which the predictor variables are usually selected and scored subjectively by expert consensus or objectively using statistical methods (den Boer et al., 2005; Higgins et al., 1997; Vincent & Moreno, 2010).

Despite their lower accuracy, scoring models are usually preferred to probability models by clinicians and health operators because they allow immediate calculation of individual patient scores as a simple sum of integer values associated with binary risk factors, without the need for any data processing system. It has also been demonstrated that in most cases, where a considerable amount of clinical information is available, their diagnostic accuracy is similar to that of probability models (Cevenini & P. Barbini, 2010, as cited in Cevenini et al., 2007). Computation facility of score models should be carefully evaluated in conjunction with their predictive performance. Too many simple models can lead to misleading estimates of a patient's clinical risk, which can be useless, counterproductive or even dangerous.

Any risk model, even if sophisticated and accurate in the local specific condition in which it was designed, loses much of its predictive power when exported to different clinical scenarios. Locally customized scoring models generally provide better performances than exported probability models. This reinforces the clinical success and effectiveness of scoring systems, the design and customisation to local conditions and/or institutions of which are usually much easier.

A limit of many scoring systems is their complex, involuted and even arbitrary design procedure that often involves contrivances to round off parameters of more sophisticated probability models to integer values. This can make their design even more complicated than that of probability models. Scoring often involves dichotomisation of continuous clinical variables to binary risk factors by identifying cut-off values from subjective clinical criteria not based on suitable optimisation techniques. However, whatever the design procedure, the main weakness of scoring models regards the interpretation of individual scores in terms of prognostic probabilities (model calibration), the reliability of which depends on the availability of a sufficient proportion of adverse outcomes and of a design procedure that provides precise individual risk estimation (Cevenini & P. Barbini, 2010). The Hosmer-Lemeshow test is commonly used to assess the calibration of probability models and therefore to manage their learning, but its results are unreliable when applied to models with discrete outputs, such as scoring systems (Finazzi et al., 2011).

This chapter provides an initial brief overview of general issues for the correct design of predictive models with binary outcomes. It broadly describes the main modelling approaches, then illustrates in more detail a method for creating score models for predicting the risk of an AHE. The method tackles and overcomes many of the above-mentioned limits. It uses a well-founded numerical bootstrap technique for appropriate statistical interpretation of simple scoring systems, and provides useful and reliable diagnostic and prognostic information (Carpenter & Bithell, 2000; DiCiccio & Efron, 1996). The whole design procedure is set out and validated by a simulation approach that mimics realistic clinical conditions. Finally, the method is applied to an actual clinical example, to predict the risk of morbidity of heart surgery patients in intensive care.

2. Model issues

Various pattern recognition approaches can be used to design models for separating and classifying patients into the two independent classes of adverse or favourable health outcome, AHE and FHE. The approaches fall into two main categories.

1. Probability models estimate a class-conditional probability, P(AHE|x), of developing the adverse outcome AHE, given a set of chosen predictor variables or features x

(Bishop, 1995; Dreiseitl & Ohno-Machado, 2002; Fukunaga, 1990; Lee, 2004). A probability threshold value, P_t, is identified for classification, over which AHE is recognized to occur, that is when P(AHE|x)>P_t; the choice of P_t depends on the clinical cost of a wrong decision and influences model classification performance (E. Barbini et al., 2007).

2. Score models evaluate risk by a discrete scale of n positive integer values s_i (i = 0, 1, 2, ..., n) which includes zero to represent null risk, but rarely provides a threshold value for classification purposes (Cevenini & P. Barbini, 2010; Vincent & Moreno, 2010).

2.1 Discrimination and calibration

Whatever the risk model, its prediction power is generally expressed by discrimination and calibration (Cook, 2008; Diamond, 1992).

Discrimination is the capacity of a classification model to correctly distinguish patients who will develop an adverse outcome from patients who will not. It must be optimized during model design by ascertaining that the model learns all the discrimination properties valid for the population, correctly from the training sample and therefore shows similar performance in different samples (generalisation ability) (Dreiseitl & Ohno-Machado, 2002; Vapnik, 1999). Though many criteria exist for evaluating model discrimination capacity (Fukunaga, 1990), sensitivity (SE) and specificity (SP), which measure the fractions of correctly classified sick and healthy patients, respectively, are commonly used for statistical evaluations of binary diagnostic test performance. SE end SP are combined in the receiver operating characteristic (ROC) curve which is a graphic representation of the relationship between the true-positive fraction (TPF = SE) and false-positive fraction (FPF = 1-SP) obtained for all possible choices of P_t. The area under the ROC curve (AUC) is the most widely used index of total discrimination capacity in medical applications (Lasko et al., 2005).

Calibration, or goodness of fit, represents the agreement between model-predicted and true probabilities of developing the adverse outcome (Hosmer & Lemeshow, 2000). Retrospective training data only provides dichotomous responses, that is presence or absence of the AHE, so true individual risk probabilities cannot intrinsically be known. The only way to derive them directly from sample data is to calculate the proportion of AHEs in groups of patients, but this obviously becomes less accurate as group size decreases. Nevertheless, from a health or clinical point of view, it is often useful to have an estimation of the level at which each event happens, using a continuous scale, such as probability. For probability models with dichotomous outcomes, calibration capacity can be evaluated by the Hosmer-Lemeshow (HL) goodness-of-fit test, based on two alternative chi-squared statistics, $\hat{H}$ and $\hat{C}$ (Hosmer & Lemeshow, 2000). The first formulation compares model-predicted and observed outcome frequencies of fixed deciles of predicted risk probability; the second compares by partitioning observations into ten groups of the same size (the last group can have a slightly different number of cases) and calculating model-predicted frequencies from average group probabilities. The $\hat{C}$-statistic is generally preferred because it avoids empty groups, although it depends heavily on sample size and grouping criterion (den Boer et al., 2005). The HL test cannot really be applied to models with discrete outputs, such as score systems, because group sizes should themselves be adjusted on the basis of discrete values (Finazzi et al., 2011).

Calibration can be improved, without changing discrimination capacity, by suitable monotonic mathematical transformations of model predicted probabilities (Harrell et al., 1996). The mean squared error between model predicted probability and observed binary outcomes is sometimes calculated as a global index of model accuracy, and has been demonstrated to incorporate both discrimination and calibration capacities (Murphy, 1973).

2.2 Generalisation, cross-validation and variable selection

Generalisation is defined as the capacity of the model to maintain the same predictive performance on data not used for training, but belonging to the same population. A high generalisation power is of primary importance for predictive models designed on a sample data set of correctly classified cases (training set). Many different procedures, which involve different correctly classified data sets for testing model performance (testing sets), have been used to control model generalisation (Bishop, 1995; Fukunaga, 1990; Vapnik, 1999). A model generalises when differences between errors of testing and training sets are not statistically significant.

Theoretically, the optimal model is the simplest possible model designed on training data and has the highest possible performance on any other equally representative set of testing data. Excessively complex models tend to overfit, i.e. give significantly lower errors on the training data than on the testing data. Overfitting produces data storage rather than learning of prediction rules. Models must be designed to avoid overfitting and improve generalisation through efficient control of the training process. This control often includes suitable techniques for the selection of predictor variables (Guyon & Elisseeff, 2003).

Computer algorithms for properly controlling overfitting are known as cross-validation or rotation techniques and make efficient use of all available data to train and test the model (Vapnik, 1999). The most common type of cross-validation procedure is k-fold, where the original sample is randomly partitioned into k subsamples, one of which is used as testing set and the other k–1 as training set. The process is then repeated k times, changing the testing set each time so that all subsamples are used for testing. A convenient variant, more appropriate in dichotomous classification, selects each subsample to contain approximately the same proportion of cases in the two classes. When k is equal to sample size, n, the procedure is called leave-one-out. One case is tested at a time at each of the n training sessions using n–1 training cases. Resampling methods also exist, and include bootstrap methods that produce different data samples by randomly extracting cases with replacement from the original dataset (Chernick, 2007).

Cross-validation can be used to compare the performance of different predictive modelling procedures and, specifically, to select different sets of predictor variables with the same model. In fact, it is convenient to select the best minimum subset of predictor variables to control generalisation and to avoid information overlap due to correlation between variables. Computer-aided stepwise techniques are usually used to obtain optimal nested subsets of variables for this purpose. To train the model, a variable is entered or removed from the predictor subset on the basis of its contribution to a significant increase in discrimination performance (typically the AUC for dichotomous classification) at each step of the process. The stepwise process stops when no variable satisfies the statistical criterion for inclusion or removal (Guyon & Elisseeff, 2003).

3. Probability models

We now provide an overview of four approaches for estimating AHE risk probability: the Bayesian classification rule (Lee, 2004), k-nearest neighbour discrimination (Beyer et al., 1999), logistic regression (Dreiseitl & Ohno-Machado, 2002; Hosmer & Lemeshow, 2000), and artificial neural networks (Bishop, 1995; Dreiseitl & Ohno-Machado, 2002). Linear and quadratic discriminant analyses and related Fisher discriminant functions were not considered because they are strictly classification methods, and although they also enable easy derivation of prediction probabilities, they have been demonstrated to be equivalent to Bayesian methods (Fukunaga, 1990).

3.1 Bayesian classifiers

Bayes's rule allows the posterior conditional probability of AHEs to be predicted as follows (Lee, 2004):

$$P(AHE\,|\,x)=\frac{P(AHE)\;p(x\,|\,AHE)}{P(AHE)\;p(x\,|\,AHE)+P(FHE)\;p(x\,|\,FHE)} \tag{1}$$

where P(AHE) and P(FHE) = 1–P(AHE) are the prior probabilities of the adverse and favourable health events, respectively, p(x | AHE), and p(x | FHE) are the corresponding class-conditional probability density functions (CPDFs) of selected features x. Posterior probability of class FHE is simply P(FHE | x) = 1–P(AHE | x).

Setting the posterior class-conditional probability threshold P_t at 0.5, the Bayes decision rule gives minimum error. It amounts to assigning patients to the class with the largest posterior probability. A higher/lower value of P_t gives rise to a smaller/larger number of patients classified at risk.

Lack of knowledge about prior probability P(AHE), i.e. the prevalence of AHE, does not affect the discrimination performance of the Bayesian classifier since it can be counterbalanced by different choices of P_t. On the contrary, a reliable estimate of prognostic probability P(AHE | x) can be obtained only if all prior probabilities and CPDFs are correctly known.

Statistical assumptions are usually made about whether CPDFs have parametric or non parametric structure. In many cases they are assumed to be of the parametric Gaussian type, because this has been proven to provide good discrimination performance, especially if a subset of predictors can be optimally selected from a large set of clinically available variables (E. Barbini et al., 2007; Fukunaga, 1990).

3.2 K-nearest neighbour algorithms

The k-nearest neighbour algorithm is among the simplest non parametric methods for assigning patients based on closest training examples in the space of features x (Beyer et al., 1999). Euclidean distance is usually used to measure between-point nearness but other metrics must be introduced if non continuous variables are considered.

In our binary classification scheme, the training phase simply consists in partitioning feature space into the two regions or classes, AHE and FHE, based on the positions of training cases. Each new patient is assigned to the region in which the greatest number of its k neighbours occurs, where k is of course a positive integer.

With two classes, it is convenient to choose an odd k to avoid situations of equality. Typically, the choice of neighbourhood size depends on the type and size of the training set;

larger values of k generally reduce the effect of noise on classification at the expense of distinction between classes.

Heuristic techniques are used to obtain the optimal value of k. A common choice is to take k equal to the square root of the total number of training cases, but cross-validation methods, such as bootstrap, are often preferred.

Although k-nearest neighbour is not strictly a probability method, it has been demonstrated that the fraction of k neighbourhood training cases falling in the AHE region is a good estimate of class-conditional risk probability (Beyer et al., 1999).

3.3 Logistic regression

Logistic regression is perhaps the most popular method for estimating risk probabilities in the medical field (Hosmer & Lemeshow, 2000). Logistic regression is a variation of ordinary regression: it belongs to the family of methods called generalized linear models, which include a linear part followed by some associated function. It can be considered a predictive model to use when the dependent response variable is dichotomous and the independent predictor variables are of any type, i.e. continuous, categorical, or both. In d-dimensional feature space, the form of the model is:

$$\log\frac{P(AHE|x)}{1-P(AHE|x)}=c_0+c_1x_1+c_2x_2+...+c_dx_d \qquad (2)$$

where "log" is the natural logarithm function, x_k (k = 1, 2, …, d) the observation data set and c_k (k = 0, 1, 2, ..., d) regression coefficients estimated from training data using maximum likelihood criteria.

The inverse of eq. 2 allows the posterior probability of AHE risk, P(AHE|x), to be modelled by a continuous S-shaped curve, even if all predictor variables are categorical. The argument of the logarithm of eq. 2 defines the probability of the outcome event occurring divided by the probability of the event not occurring and is known as the odds ratio. When it is specifically associated with dichotomous predictor variables (risk factors), it is a useful measure of the relative risk due to single risk factors. The reliability of logistic regression results is affected by linear correlations and interaction effects between predictor variables, dependence between error terms, and especially outliers.

3.4 Artificial neural networks

Artificial neural networks (or simply neural networks) are mathematical models miming the physiological learning functions of the human brain. They can be designed and trained to create optimal input-output maps of any physical or statistical phenomenon, the relationships of which may even be complex or unknown. They do not require sophisticated statistical hypotheses and account for all possible interrelations between predictor variables in a natural way. In this sense, neural networks can be considered universal approximators (Bishop, 1995).

A preliminary definition of network architecture is needed and should include number of neurons, number of layers, number and type of connections among neurons, type of neuronal activation functions and so on. Learning is the trickiest phase of neural networks: it consists of estimating network parameters (connection weights and activation thresholds) iteratively from training data, to minimize error between actual and model-estimated outputs. Feed-forward neural networks can be designed to directly estimate class-conditional posterior probabilities from predictor variables, without requiring sophisticated statistical hypotheses. Their architecture can be variably complex, but should provide one output neuron with a logistic sigmoid activation function, generating an output between 0 and 1. Neural networks have been demonstrated to provide reliable estimates of class-conditional posterior probabilities, such as the AHE risk probability, P(AHE|x), that is (Bishop, 1995):

$$\begin{aligned} &P(AHE|x)=\frac{1}{1+\exp(-f)} \\ &f=b+w_1u_1+w_2u_2+...+w_nu_n \end{aligned} \tag{3}$$

where f is a linear function of n neuron inputs u_k (k = 0, 1, 2, ..., n), originating from the outputs of n preceding connected neurons, the parameters of which are connection weights, w_k, and neuron activation bias, b.

Under-learning can lead to high prediction errors, whereas over-learning can cause overfitting which produces loss of generalisation. Artificial neural network design is therefore anything but simple. Experience is necessary to manipulate heuristic procedures for suitable definition of network architecture and to correctly use iterative numerical training techniques that stop learning when the network begins to overfit.

4. Direct score model

A scoring model is a formula that assigns points based on known information, in order to predict an unknown future outcome. Many integer score systems have been designed for clinical application to critical patients. The most popular were derived from simplification of any of the above probability models by rounding their parameters to integer values. In particular, many approximate the coefficients of logistic regression models to the nearest integer values (Higgins et al., 1997). We do not dwell on the methodology of these score models here, directing readers to the specialised literature (Vincent & Moreno, 2010). Our main interest is to identify score values that give reliable probabilities of individual risk for prognostic purposes. We discuss on the design of a very simple score system that we call a “direct score model”. We also provide a correct and useful statistical interpretation of model prognostic capacity, which can easily be extended to any other score model, even more sophisticated ones (Cevenini & P. Barbini, 2010).

4.1 Model design

Only binary predictor variables (risk factors) are used in this score model. The automatic computer procedure and model training is described by the following steps:

- All quantitative predictor variables are dichotomised by ROC curve analysis, identifying cut-off values giving equal sensitivity and specificity in relation to adverse outcomes.
- Risk factors over or under the cut-off value are coded 0 or 1, depending on whether the risk of AHE decreases or increases, respectively.
- The odds ratio of each binary variable is evaluated on the basis of the corresponding confidence interval (CI) (Agresti, 1999): variables with odds ratios not significantly greater than 1 are discarded.

- A forward iterative procedure is applied to a data sample (training set) which sums selected binary variables stepwise.
- All binary factors are reconsidered at each step, so that multiple selection of one factor gives rise to a multiple integer contribution to the score.
- At each step the risk factor providing the highest increment to AUC is included.
- Training is stopped when the cumulative increment in AUC obtained in five consecutive steps is less than 1%. This rather soft stopping criterion is used instead of well-established statistical methods (Zhou et al., 2008) to avoid selecting too few predictors, which reduces the possibility of associating an effective probability of AHE with each integer score.
- A testing dataset of the same size as the training set is used to evaluate model generalisation and to guide conclusive selection of the optimal predictor set.

Backward sessions and cross-validation trials cannot be applied because the model is non-parametric. Optimal model selection is carried out by a step-by-step analysis of model prognostic and diagnostic power. At each step w, the conditional probability of the adverse outcome (prognostic risk probability), $P_w(AHE|S_k)$, associated with each kth integer score value S_k, is estimated from sample data as the ratio of adverse events to the total number of events determining a model score S_k.

The bias-corrected and accelerated bootstrap method is applied to estimate 95% CIs of $P_w(AHE|S_k)$ using 2000 bootstrapped samples. This method makes it possible to infer complex statistics that are difficult or even impossible to represent mathematically and have proven to be theoretically and practically more accurate than other bootstrap methods (Cevenini & P. Barbini, 2010; DiCiccio & Efron, 1996). By graphic inspection of results, the convenience of grouping close scores having large 95% CI because of excessively low data frequencies is considered. The model is chosen to correspond to the iteration providing the largest number of score values or classes having sufficiently narrow and separate 95% CIs with respect to the training data, and at the same time giving testing-data probabilities falling within their 95% CIs.

Once the model is created, the score, S, associated with a generic patient is simply given by:

$$S=\sum_{i=1}^{d} p_i s_i \tag{4}$$

where d is the number of predictors in the model, p_i the binary value of the ith predictor, and s_i, its model-identified associated score. Finally, model discrimination and calibration performance are compared with a logistic regression model designed on the same training data.

All statistical procedures are evaluated at a significance level of 95%.

4.2 Simulation

Many realistic simulation experiments are carried out to validate and optimise model design. Predictor variables are all taken in binary form, skipping the dichotomisation of continuous variables. In particular, we consider d dichotomised binary predictors, obtaining $n = 2^d$ different combinations of these predictors. Each combination identifies one value of a discrete variable $x_j = j/n$ ($j = 0, 1, 2, ..., n-1$) ranging from 0 to 1. In this way two different beta probability density functions can be associated with adverse and favourable outcomes.

Beta distribution is particularly suitable for representing multinomial phenomena, such as that described by the above n discrete values. In detail, we refer to the discrete probability distribution of a multinomial variable x, the probability values of which are calculated using a beta probability density function.

Figure 1 shows an example with two different choices of the beta probability density function shape parameters, α and β, to simulate healthy and sick subjects. When the class-conditional probability density functions of a two-class classification problem are known, the highest achievable discrimination level is related to the areas of overlap. The lowest error probability of classification, ε, is given by:

$$\varepsilon=\int_{-\infty}^{+\infty} \min\{P(C_1)p(x\,|\,C_1),P(C_2)p(x\,|\,C_2)\}\,dx \tag{5}$$

where $P(C_h)$ and $p(x\,|\,C_h)$ are the prior probability and the class-conditional probability density function for class C_h (h = 1, 2), respectively. Prior probability of an adverse outcome, P(AHE), is also known as prevalence, π, and prior probability of favourable outcome, P(FHE), is 1-π. Because of the discrete nature of variable x, in our simulation study, eq. 5 can be approximated as:

$$\begin{gathered}\varepsilon=\frac{1}{n}\sum_{j=0}^{n-1}\min\left[\pi\times p(x_j|\,\mathrm{AHE}),(1\text{-}\pi)\times p(x_j|\,\mathrm{FHE})\right]\\ p(x_j|\,\mathrm{AHE})=B\left(x_j,\alpha_{\mathrm{AHE}},\beta_{\mathrm{AHE}}\right)\\ p(x_j|\,\mathrm{FHE})=B\left(x_j,\alpha_{\mathrm{FHE}},\beta_{\mathrm{FHE}}\right)\end{gathered} \tag{6}$$

where α_{AHE}, β_{AHE}, α_{FHE} and β_{FHE} are the corresponding shape parameters of beta functions, $B_{AHE} = B\left(x_j,\alpha_{AHE},\beta_{AHE}\right)$ and $B_{FHE} = B\left(x_j,\alpha_{FHE},\beta_{FHE}\right)$, related to adverse and favourable outcomes, respectively.

Eq. 6 shows that ε depends on prevalence and beta parameters. At any iteration w of the above-mentioned stepwise procedure, for any kth integer value of score S_k, the simulated "true" conditional risk probability, $P_w^t(AHE\,|\,S_k)$, can be calculated using the Bayes theorem, considering AHE prevalence, π, and the class-conditional score probabilities, $P_w^t(S_k\,|\,AHE)$ and $P_w^t(S_k\,|\,FHE)$, of adverse and favourable outcomes, respectively:

$$P_w^t(\mathrm{AHE}\,|\,S_k)=\frac{\pi\,P_w^t(S_k|\,\mathrm{AHE})}{\pi\,P_w^t(S_k|\,\mathrm{AHE})+(1\text{-}\pi)\,P_w^t(S_k|\,\mathrm{FHE})} \tag{7}$$

By assuming mutually exclusive x_j events, the true class-conditional probabilities are simply obtained from the two simulated beta distributions as the sum of all the discrete probabilities corresponding to the x_j values giving the score S_k, that is:

$$\begin{gathered}P_w^t(S_k|\,\mathrm{AHE})=\frac{1}{n}\sum_{x_j\in S_k} B\left(x_j,\alpha_{\mathrm{AHE}},\beta_{\mathrm{AHE}}\right)\\ P_w^t(S_k|\,\mathrm{FHE})=\frac{1}{n}\sum_{x_j\in S_k} B\left(x_j,\alpha_{\mathrm{FHE}},\beta_{\mathrm{FHE}}\right)\end{gathered} \tag{8}$$

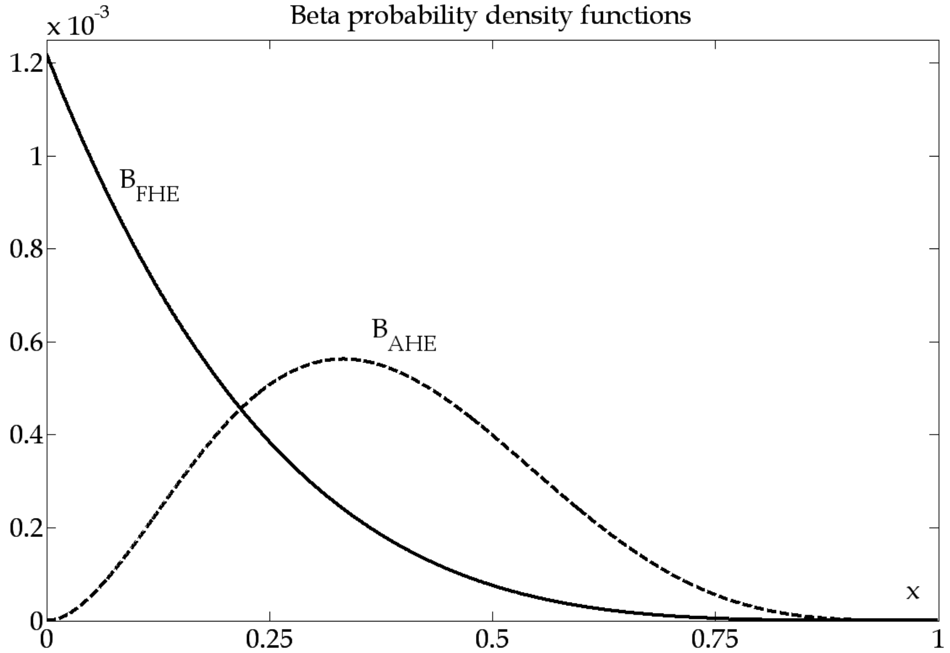


Fig. 1. Simulated probability density functions, B_{FHE} and B_{AHE}, for favourable and adverse outcomes, respectively: example with beta parameters $\alpha_{FHE} = 1$, $\alpha_{AHE} = 3$, $\beta_{AHE} = \beta_{FHE} = 5$

4.2.1 Simulation experiments

Simulation experiments are performed by randomly extracting $N = N_{AHE} + N_{FHE}$ data items from beta distributions of adverse and favourable outcomes, B_{AHE} and B_{FHE}, respectively, to form two samples of size $N_{AHE} = \pi \cdot N$ and $N_{FHE} = (1-\pi) \cdot N$. Each extracted item x_j ($j = 1, 2, ..., N$) is represented as a d-dimensional point in the discrete space of binary variables.

We use d = 12 binary variables and simulate nine different conditions corresponding to the combinations of three prevalence values and three levels of separation between event classes, obtained by changing the parameters of beta distributions. Low, medium and high separation between AHEs and FHEs are reproduced by increasing only the values of parameter α_{AHE}, specifically equal to 2, 3 and 5, respectively. The other three beta parameters are kept constant at $\alpha_{FHE} = 1$, $\beta_{AHE} = \beta_{FHE} = 5$. Prevalence values of 5%, 20% and 40% are tried. For each condition, six samples with progressively doubled sizes, namely N = 250, 500, 1000, 2000, 4000 and 8000, are extracted for a total of 54 simulation experiments covering a wide range of actual clinical situations (see also Table 1). Training data is not used because the simulation process enables the true probabilities, described above, to be evaluated exactly.

All computations are performed using MATLAB code.

4.2.2 Simulation results

The method is illustrated in detail by describing the results of a simulation of the 54 experiments performed. The experiment corresponding to N = 1000, π = 20% and $\alpha_{AHE} = 3$ is illustrated, because it is similar to an actual clinical condition that will be shown below.

Figure 2 shows the AUC values obtained using the forward selection of model features from simulated training data described above. The stopping criterion arrested the stepwise algorithm at the eleventh step, after 5 out of 12 predictor variables had been selected. In fact, the cumulative increment in AUC was about 0.8% in the last five steps (nos. 7-11). The variables are numbered in order of decreasing discrimination power. The most discriminating variable, no. 1, was entered five times ($s_1 = 5$) in the model, variable no. 2 three times ($s_2 = 3$) and variables nos. 3-5 only once each ($s_{3\text{-}5} = 1$).

Figure 3 shows the 95% confidence interval of score-associated risk probabilities identified by the bias-corrected and accelerated bootstrap method applied to simulated sample data, from step no. 2 to step no. 9. For each integer score value, the estimated 95% CI is plotted together with the corresponding true probability of AHE (calculated from the beta distribution) and the percentage of cases. The discrimination capacity of the model can be detected at every step by observing the growth of estimated AHE probability with the score, whereas calibration is demonstrated by true risk probabilities (stars), which fall in the corresponding 95% confidence interval of the training data, with the sole exception of certain high scores, where there may be too few cases.

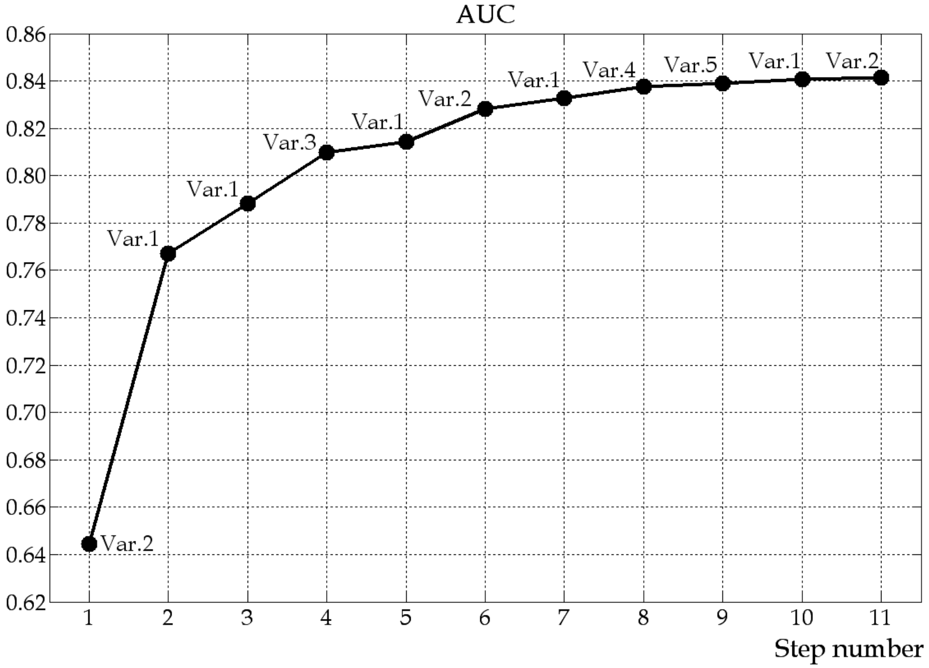


Fig. 2. Area under the ROC curve (AUC) during stepwise selection of model features from simulated data. The predictor variables entered are also indicated

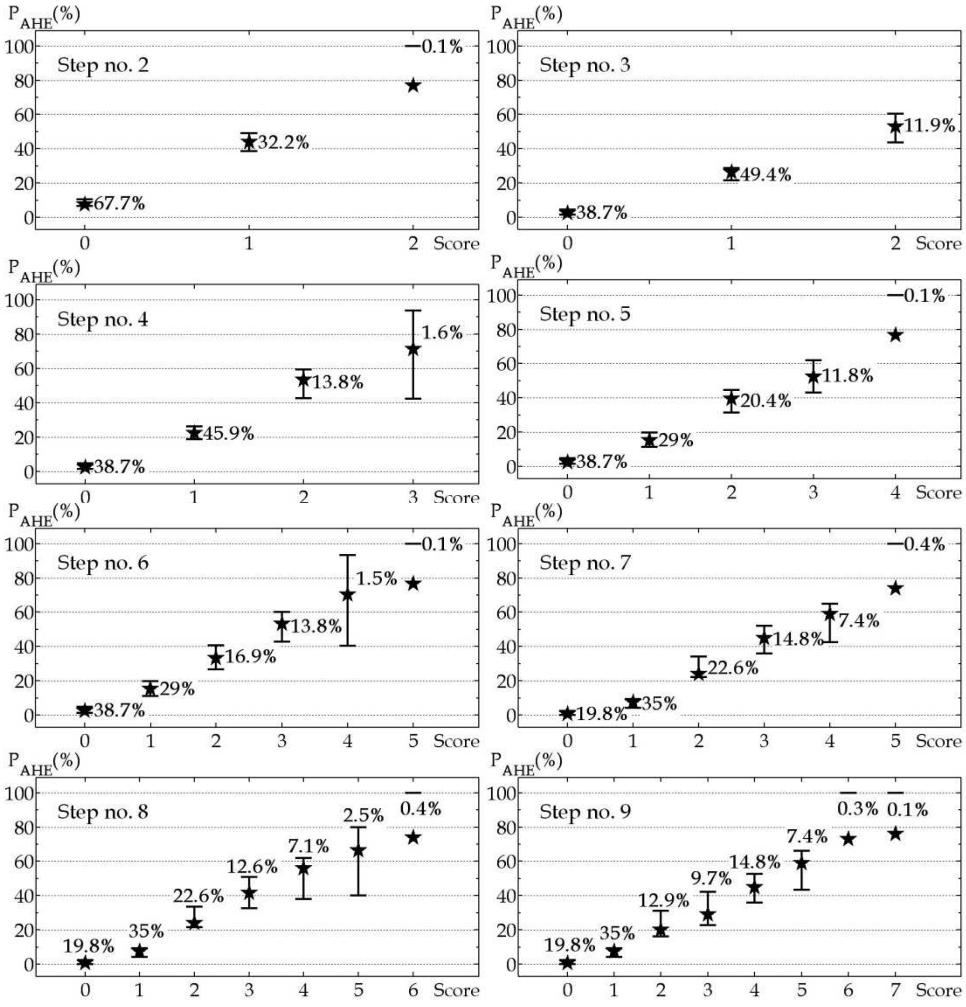


Fig. 3. 95% confidence intervals of AHE score probability, estimated from simulated training data, percentages of score cases and true probabilities (stars)

Now it is necessary to identify a model that reconciles calibration and discrimination. Excessively simple score models (few steps) have low discrimination power (low AUC) and give inopportunely separated 95% CIs. This can be observed at steps nos. 2 and 3 of Fig. 3 where only three score values (0, 1 and 2) are obtained: CIs between scores have very large gaps, suggesting that finer partitioning of the score axis can be achieved with a larger number of steps. Figure 2 indicates poor discrimination capacity of the scoring system at these initial steps.

On the contrary, if too many scores are used, as in steps nos. 7-9, the CIs are either too wide or overlap, worsening calibration accuracy. The width of score CIs increases significantly with decreasing observed frequency. For example, at step no. 4, score of 3 has only 16 cases (1.6%) and the corresponding 95% CI is so large that it completely overlaps with the previous score of 2. When the number of cases is even lower, as in step no. 9, where the highest scores of 6 and 7 have four and one cases, respectively, the bootstrap method fails to correctly estimate the CIs and the corresponding scores are totally unreliable in prognostic terms. Hence the need to combine neighbouring scores with too few cases. It is particularly convenient to pool the highest scores, which often have few cases, into a single class having a sufficient data frequency to significantly narrow the 95% CIs. For example, at step no. 6 it is useful to pool the last two scores of 4 and 5 into a single class. The pooling of adjacent scores with small data frequency enhances model prognostic reliability, usually with an insignificant reduction in discrimination capacity.

From the simulated experiment of Fig. 3, five score classes were identified as a suitable compromise between calibration and discrimination. At any step from no. 6 to no. 9, it is worthwhile combining scores greater than or equal to 4 and leaving the lower scores of 0-3 ungrouped, so as to form five score classes: 0, 1, 2, 3 and ≥ 4.

Figure 4 shows the results of pooling the three highest scores of step no. 8, which is preferred to the previous steps no. 6 and no. 7, because besides having higher discrimination capacity, the pooled class contains a greater number of cases, which narrows the related 95% CI to a greater extent. Just a small gap and a slight overlap can be observed in Fig. 4 between scores of 1 and 2, and between scores of 3 and the class of scores ≥ 4, respectively. Step no. 9 and subsequent steps not reported in Fig. 3 are discarded because no improvement can be obtained with respect to step no. 8 and CI overlap increases. Indeed, to improve the accuracy of estimates of individual probability of AHE, it could be worthwhile increasing the number of classes, tolerating a greater CI overlap. This can be done by analysing and selecting a step beyond the eighth, where the observed frequency in each class is of course significantly reduced, especially for high scores.

Comparison of the results of the three-step model with those of the eight-step pooled model shown in Fig. 4 indicates that the scoring system with five classes effectively fills the gaps between adjacent CIs of the simpler score model. At step no. 8, pooling of the highest scores does not significantly influence the discrimination capacity of the scoring system: the estimated AUC decreases slightly from 0.838 (95% CI, 0.781-0.885) to 0.827 (95% CI, 0.777-0.869).

Stepwise logistic regression applied to the training data used for the simulation example, set at statistical significance levels of 95% and 90% to enter and remove variables, respectively, selected the first five binary variables. Figure 5 compares ROC curves of the logistic model and the score model of Fig. 4. The ROC curve of true probability values, calculated from training data using beta distributions and the Bayes theorem, is also plotted (dashed line). AUCs of true data and the logistic model were 0.845 and 0.849, respectively.

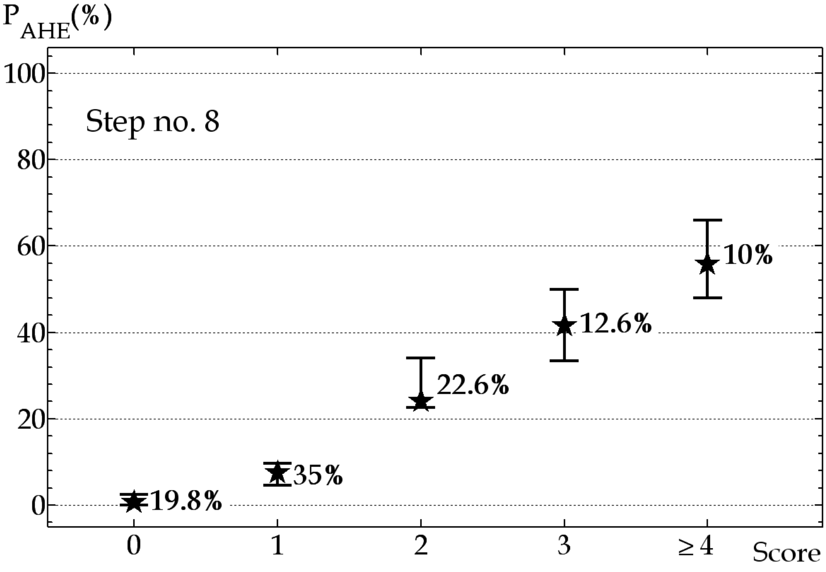


Fig. 4. 95% confidence intervals of AHE score probabilities estimated from simulated training data, percentages of score cases and true probabilities (stars) for the model identified at step no. 8

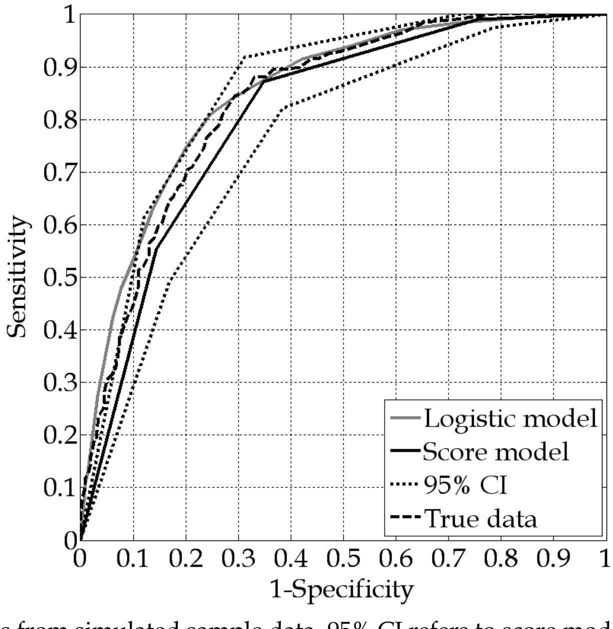


Fig. 5. ROC curves from simulated sample data. 95% CI refers to score model

When comparing model discrimination power by AUC, we have to consider that the ROC curve of the score model (continuous line) is drawn by connecting only 5 discrete points (score classes), whereas the logistic model curve (gray line) is based on more probability values. Figure 5 shows that the score model is close enough to the logistic and true ROC curves. Clearly, discretisation leads to a lower AUC, resulting in underestimation of score-model discrimination capacity. In addition, the true and logistic curves are to a large extent within the 95% CI of the score curve. Finally, in real clinical applications, logistic regression often includes continuous variables that may improve discrimination performance.

The HL goodness-of-fit test (Hosmer & Lemeshow, 2000) showed good calibration performance of the logistic model (p = 0.751). However, 95% of the training-data errors between model-estimated and true percentage risk probabilities were from about -10.5% (underestimation) and +12.0% (overestimation), revealing similar uncertainty to that of the score model.

Table 1 gives the number of score values or classes identified by the same procedure, for each of the 54 simulation experiments. It shows that the number of score classes increases with increasing sample size, prevalence and separation between event classes (decreasing error ε). The importance of estimating uncertainty suggests to keep 95% CIs of between-class probabilities separate, or slightly overlapping. This limits the identifiable number of score classes and provides reliable probability estimates. Enlargement and overlapping of 95% CIs and consequent loss of prognostic probability information depends heavily on the data frequency of score values or classes and their rate of AHEs influenced by prevalence. Small samples and/or low prevalence make it necessary to pool neighbouring scores to form classes with a sufficient number of cases to ensure a reliable estimate (narrow CI) of class probabilities.

	Low separation $\alpha_{AHE} = 2$			Medium separation $\alpha_{AHE} = 3$			High separation $\alpha_{AHE} = 5$		
Π%	5	20	40	5	20	40	5	20	40
ε% N	5.0	20.0	32.9	5.0	17.7	23.9	4.6	11.4	13.7
250	2	3	3	3	4	4	3	4	4
500	3	4	4	4	5	5	4	5	5
1000	4	4	4	4	5	5	4	5	6
2000	4	5	5	5	5	6	5	6	6
4000	5	5	5	5	6	6	5	6	6
8000	5	6	6	6	6	7	6	7	7

Table 1. Simulation experiments: largest number of score classes having sufficiently narrow and separate 95% confidence intervals of prognostic probability. αAHE = shape parameter of AHE beta distribution; Π = prevalence; ε = lowest error probability of classification; N = sample size

Simulation experiments suggests grouping scores into classes when frequencies are less than about 3% and 10% of the whole sample for N = 8000 and N = 250, respectively. Only two classes are recognised in the worst condition of minimum sample size (N = 250), minimum prevalence ($\Pi = 5\%$) and low separation between health events ($\alpha_{AHE} = 2$). A maximum of seven score-classes is identified in conditions of large sample size (N = 8000), high prevalence and high separation between event classes. Although more score classes could be achieved with greater CI overlap, the cost would be unreliable estimates.

The discrimination of the different simulation experiments is assessed by AUC of true simulated probability calculated using beta functions. Conditions of large overlap between areas of beta functions ($\alpha_{AHE} = 2$) lead to values of true AUC ranging from 0.72 to 0.75; medium overlap ($\alpha_{AHE} = 3$) gives AUC values in the range 0.82-0.85 and the conditions of greatest separation ($\alpha_{AHE} = 5$) produce AUCs between 0.92 and 0.95.

4.3 Clinical example

The approach was applied to actual clinical data of critical patients in the intensive care unit to evaluate their risk of morbidity after heart surgery.

We used a sample of 1040 adult patients younger than 80 years, who underwent coronary artery bypass grafting and were admitted to the intensive care unit of the Department of Surgery and Bioengineering of Siena University. 212 patients developed at least one serious postoperative complication (cardiovascular, respiratory, neurological, renal, infectious or hemorrhagic), corresponding to a morbidity of 20.4% (Cevenini & P. Barbini, 2010, as cited in Cevenini et al., 2007). The data was split randomly into a training and a testing set of the same size (520 cases), with the same number of patients with morbid conditions in each set (106 cases) to avoid misleading bias in the results.

Table 2 describes the 15 clinical variables used for score model design, six of which were binary in origin. The other nine continuous variables were dichotomised using cut-off values associated with the point of equal sensitivity and specificity on the respective ROC curves. Three of the resulting 15 binary variables were discarded because their odds ratios of morbidity were not significantly greater than 1. This left a total of 12 variables for training the score model, as in the simulation experiments.

This real clinical situation was similar to the simulation experiment with N = 500 and $\Pi = 20\%$ (see Table 1). Consulting Table 1, we expected to develop a score model with 4 or 5 classes, depending on the level of data separation between normal and morbid patients.

Figure 6 shows the stepwise procedure used to select the model variables. After step no. 8, AUC values of testing data (dashed line with stars) decreased and diverged from training data AUCs (continuous line with dots). This indicated overfitting that was possible because the criterion used to stop the training procedure was deliberately soft, to allow inclusion of more steps than needed for generalisation. In fact, as previously illustrated in the simulation results, investigation of extra steps can be useful to optimise model prognostic power through score pooling. Steps nos. 6, 7 and 8 gave similar prognostic performance, so we chose step no. 8, thus obtaining higher discrimination (greater AUC). A convenient class was formed by pooling scores greater than 3, as shown in Fig. 7. All 95% CIs of adjacent scores or classes were well-separated and all testing score probabilities (stars) fell within their corresponding CIs, thereby ensuring high prognostic reliability of the model. The pooling of the highest scores of the eight-step model led to a slight but not statistically significant reduction in discrimination performance: the estimated training and testing AUCs decreased from 0.851 (95% CI, 0.781-0.909) to 0.835 (95% CI, 0.764-0.895) and from 0.841 (95% CI, 0.775-0.900) to 0.816 (95% CI, 0.743-0.879), respectively.

Variable description	**Acronym**	**Type**	**Cut-off**	**Steps**
Inotropic heart drugs	IHD	Binary		1,4,10 (LR)
O_2 delivery index	DO_2I	Continuous	< 280 ml/min/m^2	2 (LR)
Peripheral vascular disease	PVD	Binary		3,9 (LR)
O_2 extraction ratio	O_2ER	Continuous	≥ 38%	5 (LR)
Emergency	EM	Binary		6
CO_2 production	VCO_2	Continuous	< 180 ml/min	7
Pulmonary artery hypertension	PAH	Binary		8 (LR)
Cardio-pulmonary bypass time	CPB	Continuous	≥ 2 hours	11 (LR)
Intra aortic balloon pump	IABP	Binary		12 (LR)
Creatinine	Cr	Continuous	≥ 1 mg/l	NE (LR)
Potassium	K	Continuous	≥ 4.1 mEq/l	NE (LR)
Haemoglobin	Hb	Continuous	< 9.6 g/dl	NE
Cardiac index	CI	Continuous	< 2.4 l/min/m^2	NS (LR)
Mean arterial pressure	MAP	Continuous	> 95 mmHg	NS
Previous heart surgery	Re-do	Binary		NS

Table 2. Clinical variables, cut-off values for the dichotomisation of continuous variables and score-model entry steps. NE = not entered; NS = not statistically significant; LR = variable selected by stepwise logistic regression

Two logistic regression models were designed to compare the score model results on the same training data with the 15 clinical variables of Table 2. The first model, named LogCV, used the original continuous variables and the second (LogBV) dichotomised them (see Table 2). The stepwise regression procedure selected ten clinical variables (see Table 2) and provided training-data AUC values of 0.906 (HL test, p = 0.135) and 0.871 (HL test, p = 0.557) for LogCV and LogBV, respectively. Figure 8 compares the ROC curves. The LogCV ROC curve (continuous gray line) showed the greatest discrimination performance, mainly because the model selected many continuous variables (6 out of 10). Except for the highest specificity values, where the discretisation effect of scoring was more evident, the score model ROC curve (continuous black line) did not differ significantly from that of LogBV (dashed gray line), which was inside the respective 95% CI and close enough to the score-model points. Model scores computed using the testing data gave a ROC curve (dashed black line) not significantly different from the training data curve. Finally, it should be noted that the discrimination performance of logistic models decreased considerably when applied to testing data (ROC curves not reported in Fig. 8): AUCs of logCV and logBV were reduced to 0.879 and 0.826, respectively, thus suggesting a possible overfitting.

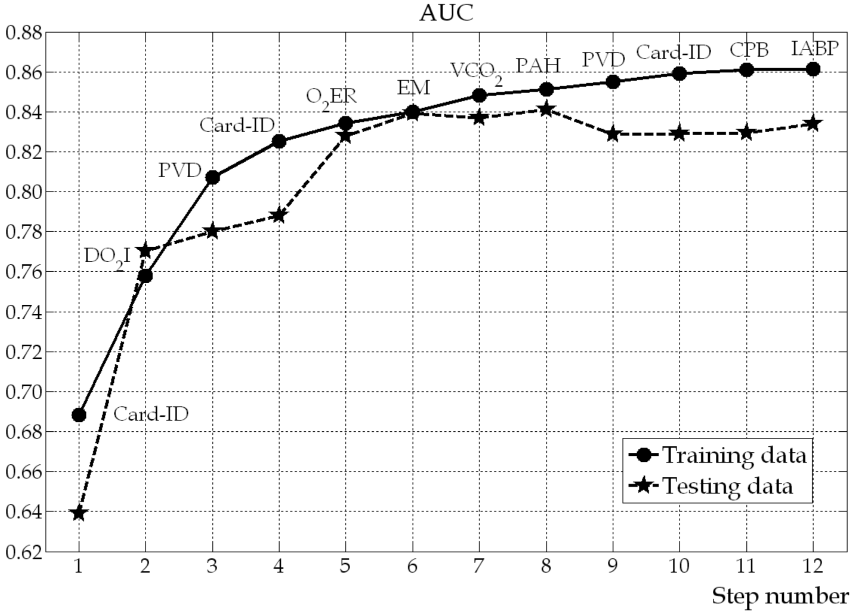


Fig. 6. Area under the ROC curve (AUC) during the stepwise selection of model features from clinical data. The predictor variables entered are also indicated

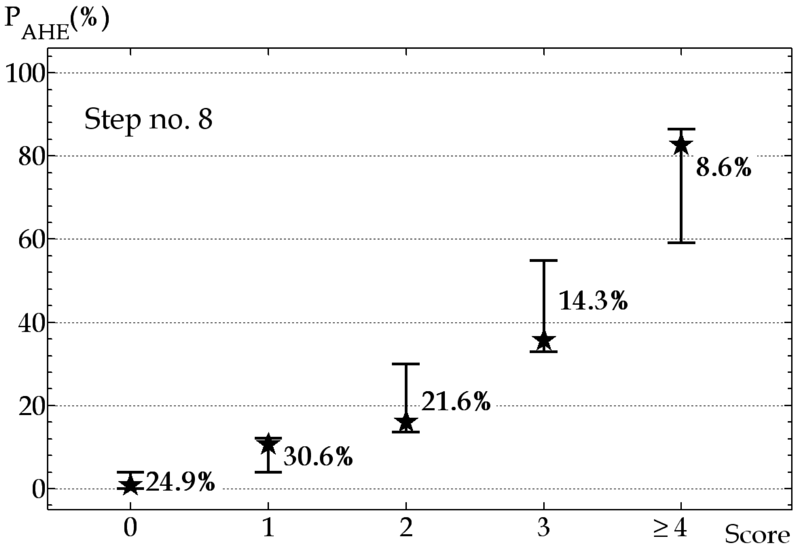


Fig. 7. Estimated 95% confidence intervals of AHE score probabilities from clinical training data, percentages of score cases and testing-data probabilities (stars) for the eight-step model chosen

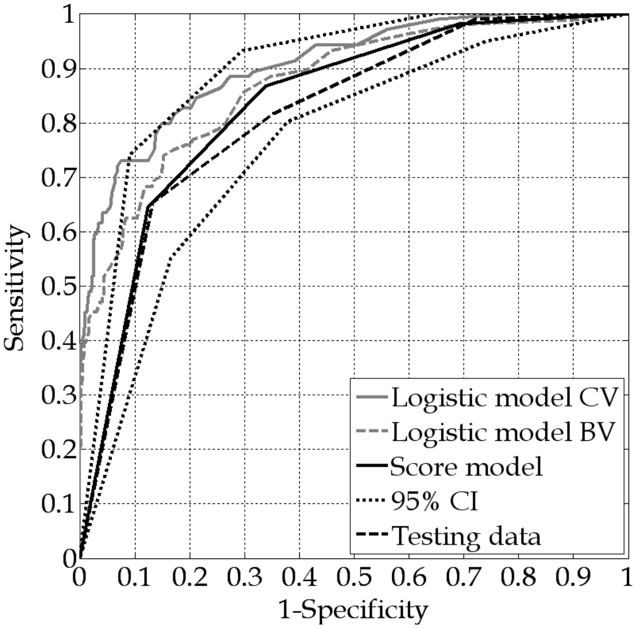


Fig. 8. ROC curves from clinical data. 95% CI refers to the score model. CV = also with continuous variables; BV = with binary variables only

5. Discussion

Many quantitative methods for assessing the health risk of critical patients have been developed in past and recent literature (E. Barbini et al., 2007; den Boer et al., 2005; Vincent & Moreno, 2010). They aim to provide objective and accurate information about patient diagnosis and prognosis. Experience has shown that simplicity of use and effectiveness of implementation are the most important requirements for their success in routine clinical practice. Scoring systems respond well to these requirements because their outcomes are accessible in real time without the use of advanced computational tools, thus allowing decisions to be made quickly and effectively. Many clinical applications can profit from their simplicity. For example, they are often used to suggest alternative treatments and organize intensive care resources, where surveillance of vital functions is the primary goal.

Other important benefits of score models are their easy updating and customisation to local institutions. In fact, because the standardisation of local practices is difficult and patient populations may differ, it is now accepted that predictive models must be locally validated, tuned and periodically updated to provide correct risk-adjusted outcomes. All models suffer from the limitation of foreseeing better future treatments and improving prognosis (den Boer et al., 2005). Even very accurate predictive models, when exported to clinical contexts different from those in which they were designed, have often proved unreliable (Murphy-Filkins et al., 1996). Appropriate design and local customisation of excessively sophisticated models is often easier said than done, especially in health centres where there is little technical expertise in developing models that can generalise, i.e. preserve their predictive performance on future data. On the contrary, simple score models can easily and frequently be updated to learn from new correctly-classified cases and are quite tolerant to missing data. This is very useful in clinical practice where data is usually scarce and training on as much available data as possible is of fundamental importance (Cevenini & P. Barbini, 2010, as cited in P. Barbini et al., 2007).

A major problem with score models is that they are difficult to calibrate, i.e. associate reliable estimates of prognostic risk probability with each score. Nevertheless, correct estimation of individual probability of adverse outcome for hospitalized critical patients is useful for prevention, treatment and quantification of health problems and costs. It can help experienced physicians to improve clinical management by optimizing the monitoring of patient status and enhancing the quality of care, and allow new generations of doctors to be better trained during postgraduate specialization and internship. Moreover, reliable knowledge of risk factors and their impact on clinical course and future quality of life can encourage public health policy for risk reduction (Hodgman, 2008).

The proposed method offers a simple risk-assessment system that associates a reliable estimate of the individual probability of developing an adverse event with predicted scores. The model is a very simple score of risk factors chosen, one or more times, by a stepwise procedure based on maximising discrimination through ROC analysis. No hypotheses or statistical models are involved. Since conventional methods for evaluating calibration, such as the Hosmer-Lemeshow test (Hosmer & Lemeshow, 2000), are unreliable for scoring systems, we analysed the 95% confidence interval of sample-estimated risk probabilities associated with each score step by step. The experimental score probability is easily evaluated by calculating the sampling rate of adverse outcomes having that score.

Unfortunately, the statistics of the sampling error are not simple to derive. We therefore preferred to use bootstrap resampling, a method commonly used in statistical inference to estimate confidence intervals (Carpenter & Bithell, 2000; DiCiccio & Efron, 1996). The bootstrap method is simpler and more general than conventional approaches; it requires no great expertise in mathematics or probability theory and is based on assumptions that are less restrictive and easier to control. The method can be used to evaluate statistics that are difficult or impossible to determine by conventional methods. We used an elaboration of the simplest bootstrap method of percentile intervals, known as bias–corrected and accelerated intervals, which avoids estimate bias and offers substantial advantages over other bootstrap methods, both in theory and practice (Chernick, 2007). Our simulation experiments confirmed the method's accuracy in estimating 95% CI of prognostic probabilities: when true probabilities were related to score values, or classes, with a sufficient number of sampled training data, they always fell within bootstrap-estimated 95% CIs (see Fig. 3). Bootstrap techniques are not too complex in a clinical environment, since nowadays many available packages for data processing include them for calculating confidence intervals. In any case, they are used exclusively during model design.

As shown in Fig. 3, step by step graphical inspection of probability CIs made it possible to choose the best model to compromise between calibration and discrimination, also suggesting convenient pooling of adjacent scores that gave large and overlapping CIs due to an insufficient number of cases or adverse events. The controlled simulation experiments showed that good calibration was achieved with a limited number of score classes, up to a maximum of seven in experiments with the biggest sample size, and high prevalence and separation between event classes (see Table 1). More classes could be identified if greater overlap of close scores were allowed, but when the number of classes became excessive, there were problems of overfitting. We also saw that a logistic model designed on the same training data provided nearly continuous probability estimates, the uncertainty of which was similar to that achieved by the score model. Significant improvement of discrimination performance could only be appreciated when continuous variables were also included in the logistic model, as in the clinical example described. This analysis can enable medical staff to select the best scoring system for any specific clinical context.

6. Conclusion

In critical care medicine, scoring systems are often designed exclusively on the basis of discrimination and generalisation characteristics (diagnostic capacity), at the expense of reliable individual probabilities (prognostic capacity). Our proposed approach that weighs both these capacities is validated by suitable simulation experiments, which also allow design conditions and application limits of scoring systems to be investigated for correct prediction of critical patient risk in a real clinical context.

The bias-corrected and accelerated bootstrap method for evaluating the 95% confidence interval, CI, of individual prognostic probabilities provides reliable estimates of true simulated probabilities. CIs are calculated for each score and at each step of scoring-system design. By increasing the number of steps, model discrimination power (greater AUC) and prognostic information (greater number of different score values) increases but widening and overlap of 95% CIs soon occurs, so that it becomes convenient to pool adjacent scores into score classes. The maximum number of different score classes giving distinct prognostic information, that is having narrow and less overlapping 95% CIs, increases with increasing sample size and prevalence of adverse outcome and decreasing error probability of classification. It is strongly limited by reduced frequency of score cases and the respective rate of adverse events: in our simulated experiments, which covered a wide range of real conditions, it varied from 2 to 7.

Application of the method to a real clinical situation demonstrated that the technique can be a simple practical tool, providing useful additional prognostic information to associate with classes of scores, and enabling doctors to choose the best risk score model to use in their specific clinical context.

7. Acknowledgment

This work was partly financed by the University of Siena, Italy.

8. References

Agresti, A. (1999). On Logit Confidence Intervals for the Odds Ratio with Small Samples. *Biometrics,* Vol.55, pp. 597-602, ISSN 0006-341X

Barbini, E.; Cevenini, G.; Scolletta, S.; Biagioli, B.; Giomarelli, P. & Barbini, P. (2007). A Comparative Analysis of Predictive Models of Morbidity in Intensive Care Unit after Cardiac Surgery – Part I: Model Planning. *BMC Medical Informatics and Decision Making,* Vol.7, No.35, (22 November 2007), pp. 1-16, ISSN 1472-6947, Available from http://www.biomedcentral.com/1472-6947/7/35

Beyer, K.; Goldstein, J.; Ramakrishnan, R. & Shaft, U. (1999). When is Nearest Neighbor Meaningful?, *Proceedings of the 7th International Conference on Database Theory*, pp. 217-235, ISBN 3-540-65452-6, Jerusalem, Israel, January 10-12, 1999

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition,* Oxford University Press, ISBN 0-19-853864-2, Oxford, UK

Carpenter, J. & Bithell, J. (2000). Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians. *Statistics in Medicine,* Vol.19, No.9, pp. 1141-1164, ISSN 0277-6715

Chernick, M.R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers*, Wiley, ISBN 978-0-471-75621-7, New York, USA

Cevenini, G. & Barbini, P. (2010). A Bootstrap Approach for Assessing the Uncertainty of Outcome Probabilities when Using a Scoring System. *BMC Medical Informatics and Decision Making,* Vol.10, No.45, (26 August 2010), pp. 1-9, ISSN 1472-6947, Available from http://www.biomedcentral.com/1472-6947/10/45

Cook, N.R. (2008). Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve. *Clinical Chemistry*, Vol.54, pp. 17-23, ISSN 1339-1348, Available from http://www.clinchem.org/cgi/content/full/54/1/17

den Boer, S.; de Keizer, N.F. & de Jonge, E. (2005). Performance of Prognostic Models in Critically Ill Cancer Patients – A Review. *Critical Care,* Vol.9, pp. R458-R463, (8 July 2005), ISSN 1364-8535, Available from http://ccforum.com/content/9/4/R458

Diamond, G.A. (1992). What Price Perfection? Calibration and Discrimination of Clinical Prediction Models. *Journal of Clinical Epidemiology,* Vol.45, No.1, pp. 85-89, ISSN 0895-4356

DiCiccio, T.J. & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, Vol.11, pp. 189-228, ISSN 0883-4237

Dreiseitl, S. & Ohno-Machado, L. (2002). Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. *Journal of Biomedical Informatics*, Vol.35, no.5-6, pp. 352-359, ISSN 1532-0464

Finazzi, S.; Poole, D.; Luciani, D.; Cogo, P.E. & Bertolini, G. (2011). Calibration Belt for Quality-of-Care Assessment Based on Dichotomous Outcomes. *PLoS One,* Vol.6, No.2, (23 February 2011), e16110, ISSN 1932-6203, Available from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0016110

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, Academic Press, ISBN 978-0-12-269851-4, Boston, USA

Guyon, I. & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, Vol.3, No.7-8, pp. 1157-1182, ISSN 1532-4435

Harrell, F.E. Jr; Lee, K.L. & Mark, D.B. (1996), Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine,* Vol.15, No.4, pp. 361-387, ISSN 0277-6715

Higgins, T.L.; Estafanous, F.G.; Loop, F.D.; Beck, G.J.; Lee, J.C.; Starr, N.J.; Knaus, W.A. & Cosgrove III, D.M. (1997). ICU Admission Score for Predicting Morbidity and Mortality Risk after Coronary Artery Bypass Grafting. *The Annals of Thoracic Surgery*, Vol.64, No.4, pp. 1050-1058, ISSN 0003-4975

Hodgman, S.B. (2008). Predictive Modeling & Outcomes. *Professional Case Management,* Vol.13, pp. 19-23, ISSN 1932-8087

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression,* Wiley, ISBN 0-4716-1553-6, New York, USA

Lasko, T.A.; Bhagwat, J.G.; Zou, K.H. & Ohno-Machado, L. (2005). The Use of Receiver Operating Characteristic Curves in Biomedical Informatics. *Journal of Biomedical Informatics*, Vol.38, No.5, pp. 404-415, ISSN 1532-0464

Lee, P.M. (2004). *Bayesian Statistics - An Introduction,* Arnold, ISBN 0-340-81405-5, London, UK

Marshall, G.; Shroyer, A.L.W.; Grover, F.L. & Hammermeister K.E. (1994). Bayesian-Logit Model for Risk Assessment in Coronary Artery Bypass Grafting. *The Annals of Thoracic Surgery*, Vol.57, No.6, pp. 1492-1500, ISSN 0003-4975

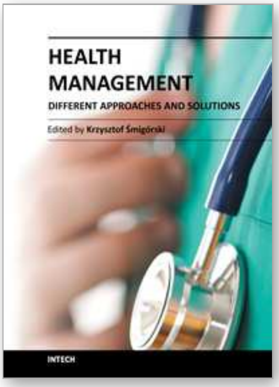
Murphy, A.H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, Vol.12, No.4, pp. 595-600, ISSN 0021-8952, Available from http:// journals. ametsoc.org/toc/jam/12/4

Murphy-Filkins, R.; Teres, D.; Lemeshow, S. & Hosmer, D.W. (1996). Effect of Changing Patient Mix on the Performance of an Intensive Care Unit Severity-of-Illness Model: How to Distinguish a General from a Specialty Intensive Care Unit. *Critical Care Medicine,* Vol.24, No.12, pp. 1968-1973, ISSN 0090-3493

Vapnik, V.N. (1999). *The Nature of Statistical Learning Theory*, Springer-Verlag, ISBN 0-387-98780-0, New York, USA

Vincent, J.L. & Moreno, R. (2010). Clinical Review: Scoring Systems in the Critically Ill. *Critical Care,* Vol.14, No.2 (207), pp. 1-9, ISSN 1364-8535

Zhou, X.H.; Li, S.M. & Gatsonis, C.A. (2008). Wilcoxon-Based Group Sequential Designs for Comparison of Areas Under Two Correlated ROC Curves. *Statistics in Medicine,* Vol.27, No.2, pp. 213-223, ISSN 0277-6715



Health Management - Different Approaches and Solutions
Edited by Dr. Krzysztof Smigorski

ISBN 978-953-307-296-8
Hard cover, 498 pages
Publisher InTech
Published online 14, December, 2011
Published in print edition December, 2011

The development in our understanding of health management ensures unprecedented possibilities in terms of explaining the causes of diseases and effective treatment. However, increased capabilities create new issues. Both, researchers and clinicians, as well as managers of healthcare units face new challenges: increasing validity and reliability of clinical trials, effectively distributing medical products, managing hospitals and clinics flexibly, and managing treatment processes efficiently. The aim of this book is to present issues relating to health management in a way that would be satisfying for academicians and practitioners. It is designed to be a forum for the experts in the thematic area to exchange viewpoints, and to present health management's state-of-art as a scientific and professional domain.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Paolo Barbini and Gabriele Cevenini (2011). Design of Scoring Models for Trustworthy Risk Prediction in Critical Patients, Health Management - Different Approaches and Solutions, Dr. Krzysztof Smigorski (Ed.), ISBN: 978-953-307-296-8, InTech, Available from: http://www.intechopen.com/books/health-management-different-approaches-and-solutions/design-of-scoring-models-for-trustworthy-risk-prediction-in-critical-patients

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821