A new approach to categorising continuous variables in prediction models: proposal and validation

Irantzu Barrio [‡1,4] Inmaculada Arostegui[1,4,5] María-Xosé Rodríguez-Álvarez[2]
José-María Quintana[3,4]

[1] Departamento de Matemática Aplicada, Estadística e Investigación Operativa,
Universidad del País Vasco UPV/EHU

[2] Departamento de Estadística e Investigación Operativa. Universidade de Vigo

[3] Unidad de Investigación, Hospital Galdakao-Usansolo

[4] Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC)

[5] BCAM - Basque Center for Applied Mathematics

Abstract

When developing prediction models for application in clinical practice, health practitioners usually categorise clinical variables that are continuous in nature. Although categorisation is not regarded as advisable from a statistical point of view, due to loss of information and power, it is a common practice in medical research. Consequently, providing researchers with a useful and valid categorisation method could be a relevant issue when developing prediction models. Without recommending categorisation of continuous predictors, our aim is to propose a valid way to do it whenever it is considered necessary by clinical researchers. This paper focuses on categorising a continuous predictor within a logistic regression model, in such a way that the best discriminative ability is obtained in terms of the highest area under the receiver operating characteristic curve (AUC). The proposed methodology is validated when the optimal cut points' location is known in theory or in practice. In addition, the proposed method is applied to a real data set of patients with an exacerbation of chronic obstructive pulmonary disease, in the context of the IRYSS-COPD study where a clinical prediction rule for severe evolution was being developed. The clinical variable PCO_2 was categorised in a univariable and a multivariable setting.

[‡]Corresponding author: E-mail: irantzu.barrio@ehu.eus, Tel.: +34-946012504, Address: Departamento de Matemática Aplicada, Estadística e Investigación Operativa. Universidad del País Vasco UPV/EHU

Keywords: categorisation; prediction models; cut point; validation

1 Introduction

Prediction models are currently relevant in a number of fields, including medicine. Decisions such as the most appropriate treatment for a disease, or whether or not a given patient should be discharged, etc., are based on the individual patient's risk of suffering some unfavourable event, and such a risk is often measured on the basis of clinical variables that are continuous in nature.

When developing prediction models, the selection of the predictors or covariates (clinical variables) to be used in the model is essential. From a statistical point of view, categorising continuous variables is not regarded as advisable, since it may entail a loss of information and power [1,2]. Additionally, there are statistical modelling techniques such as generalised additive models (GAM) [3,4], which do not require any assumption of linearity between predictors and response variables, and so allow for the relationship between the predictor and the outcome to be modelled more appropriately. Yet in clinical research and, more specifically, in the development of prediction models for use in clinical practice, both clinicians and health managers call for the categorisation of continuous variables. Indeed, in a recent survey of the epidemiological literature, in the 86% of the papers included in the study, the primary continuous predictor was categorised, of which the 78% used 3 to 5 categories [5]. In our opinion, there are several reasons for this. Firstly, in clinical practice, the application of results obtained from techniques such as GAM is not always viable. It requires specific software which is not always possible to use at consulting rooms or emergency departments. On the other hand, decisions in clinical practice are often taken on the basis of an individual patient's risk level, which is strongly related to a categorisation of that patient's clinical variables. Yet, despite the fact that categorisation is a common practice in clinical research, there are no unified criteria for categorising continuous variables. Indeed, categorisation is very often based on percentiles, even though this is known to have drawbacks [6]. Moreover, even when categorisation is based on clinical criteria, it has been shown that it can vary enormously from one practitioner or hospital (or even country) to another. For instance, a meta-analysis conducted by Lim and Kelly [7] showed that reported cut-off values for partial pressure of carbon dioxide in the blood (PCO_2) for hypercapnia screening ranged from 30 to 46 mnHg.

Previous work has been done on the categorisation of continuous variables. A review of these methods shows that these have been based: firstly, on the graphical relationship between the predictor and the outcome; and secondly, on the minimum p-value approach [8]. Moreover the aim in almost all cases has been to seek a single cut point, or, expressed in another way, to dichotomise the continuous predictor [9–11]. However, the use of more than two categories may be preferable, since this serves to reduce the loss of information and enables the relationship between the covariate and the response variable to be retained. In the context where the outcome of interest takes only two possible values, the search for more than one cut point has been considered for instance by Tsuruta and Bax[12] and Barrio et al[13]. Tsuruta and Bax propose a parametric method for obtaining cut points based on the overall discrimination C-index[14], which is equivalent to the area under the receiver operating characteristic curve (AUC). The authors showed the optimal location of cut points in a case where the distribution of the predictor variable is known, and illustrated the proposal for application to a normal distribution. Yet, in routine clinical practice and, by extension, in medical research, variables of interest do not usually respond to either a normal or a known distribution. On the other hand, Barrio et al proposed a method based on a graphical display using GAM with P-spline smoothers to determine the relationship between the continuous predictor and the outcome.

Despite the fact that both approaches have proven to be useful they suffer from the limitation of only being applicable in a univariable setting. Accordingly, we propose in this paper a new approach for the selection of optimal cut points that allows for more than one cut point to be selected as well as the possibility of being used in a multivariable setting. Specifically our study has two main aims: firstly, to propose a new approach for the selection of optimal cut points for categorising continuous variables in logistic prediction models; and secondly, to validate the proposed approach. Two different algorithms, called *AddFor* and *Genetic*, are proposed for the selection of the cut points which maximise the AUC, and the performance is evaluated and compared by means of simulations. Validation of the categorisation method is performed in two different settings: 1) under defined theoretical conditions, where the optimal cut points are known; and, 2) under empirical situations where the original variable is observed as categorical although an underlying continuous latent variable is supposed.

The rest of the paper is organised as follows. Section 2 provides a description of the IRYSS-COPD study of patients suffering from an exacerbation of chronic obstructive pulmonary disease which motivated the development of the methodology presented in this paper. Section 3 outlines the method proposed for categorising continuous variables in clinical prediction models where the response variable is dichotomous. In Section 4, the validation process is described, with a comparison of both cut point selection algorithms in different settings, namely, under theoretical and empirical defined conditions. Additionally, the results obtained from the validation study are reported. Section 5 describes the software implementation. Section 6 describes the application of the proposed methodology to the IRYSS-COPD study data set. Finally, the paper closes with a discussion in Section 7 in which the findings are reviewed and conclusions are drawn.

2 The IRYSS-COPD study

Chronic obstructive pulmonary disease (COPD) is one of the most common chronic diseases, and its prevalence is expected to increase over the next few decades[15]. COPD is a leading cause of death in developed countries, and patients with COPD generally have a substantial deterioration in their quality of life[16]. Exacerbation of COPD (eCOPD) is defined as an event in the natural course of a patient's COPD characterised by a change in baseline dyspnea, cough, and/or sputum that is beyond normal day-to-day variations and that may have warranted a change in medication or treatment[17]. Patients often experience eCOPD, and these often require assessment in an emergency department (ED) and hospitalisation. Exacerbations play a major role in the burden of COPD, its evolution, and its cost[18]. Some exacerbations are quite severe, leading to death or the need for invasive mechanical ventilation (IMV); others are more moderate, requiring little more than an adjustment of the patient's current medical therapy. Currently, ED physicians must rely largely on experience and personal criteria for gauging how an eCOPD will evolve. Accordingly, the development of clinical prediction rules in this context would be of great importance to help ED physicians to make better informed decisions about treatment.

The IRYSS-COPD study (IRYSS: Red de investigación cooperativa para la Investigación en Resultados de Salud y Servicios Sanitarios - Co-operative Health Outcomes & Health Services Research Network) was created to address gaps for identifying eCOPD patients whose clinical situation is appropriate for admission to the hospital, and to develop and validate severity scores for eCOPD exacerbations[19]. In this study, a sample of 2487 patients with eCOPD attending the EDs of 16 participating hospitals in Spain was collected. Information was recorded as follows: at the date on which patients were evaluated at the ED; at the date on which the decision was made to admit patients or discharge them home from the ED; and during follow-up after admission to the hospital or discharge home. Data collected upon arrival in the ED included socioeconomic data, information about the patient's respiratory function (arterial blood gases, respiratory rate, dyspnea), and presence of other pathologies recorded in the Charlson Comorbidity Index. The consciousness level was measured by the Glasgow coma scale which was dichotomised as follows: altered consciousness defined as a score of < 15 points, unaltered consciousness as a score of 15 points[20]. Additional data collected in the ED at the time a decision was made to admit or discharge the patient included the patient's symptoms, signs, and respiratory status at that moment.

One of the goals of the IRYSS-COPD study was to develop a clinical prediction rule for the short term very severe evolution of eCOPD defined as any of the following: death, Intensive Care Unit (ICU) admission, need for IMV, and/or cardiac arrest. After a preliminary analysis, the predictors selected to be included in the prediction model were the Glasgow Coma scale, heart rate and the arterial blood gas PCO_2. However, the covariate PCO_2 had not a linear relationship with the outcome and hence it required to be introduced either modelled with a smooth function or in a categorised version. The clinical researchers involved in the study claimed for a categorised version of this predictor, but as mentioned earlier, there was not a previously fixed cut point criteria in the literature[7]. The authors previously proposed the categorisation of the covariate PCO_2 which relayed on a graphical display[13]. However, at this time we considered developing a more general methodology to obtain optimal cut points to categorise the continuous predictor variable PCO_2, so that we could obtain the best categorised version to be introduced in the short term very severe evolution of eCOPD prediction model.

3 Methods

This section describes the methodology proposed to categorise continuous predictors in logistic regression models. Once the needed background and notation have been introduced, we describe the proposed methodology and the two algorithms for its implementation in Section 3.1. It should be noted that when developing the logistic regression model, the obtained AUC might be biased upward when the same data-set is used to fit the model and to compute the AUC. Accordingly, we considered correcting the overestimation of the AUC in the logistic model, which is explained in detail in Section 3.2. In addition to this, in clinical practice it might be needed to select which is the most desirable number of cut points. Therefore in Section 3.3 we present two possible approaches to select the best number of cut points.

Suppose one has a dichotomous response variable Y, and a continuous predictor X. Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of the outcome characteristic respectively. Then, the logistic regression model for Y is written as a linear function in the logistic transformation (*logit*) as it is shown in equation (1), where β_0 is the intercept and β_1 is the regression coefficient for X.

$$logit(P(Y=1|X)) = log\frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X. \tag{1}$$

For a binary outcome, the AUC is the most commonly used performance measure to evaluate the discriminative ability of a prediction model. More specifically, given a sample $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$, the coefficients β_0 and β_1 are estimated by maximum likelihood and an iterative weighted least squares algorithm (denote $\hat{\beta}_0$ and $\hat{\beta}_1$). More detail about estimation methods can be seen in McCullagh and Nelder[21]. Let $\hat{p}_i = logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ be the estimated probability for subject i, then the AUC is frequently estimated by the Mann-Whitney statistic[22] which is given as:

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{j \in D_{Y=0}} \sum_{m \in D_{Y=1}} I[\hat{p}_j, \hat{p}_m],$$

where $D_{Y=1}$ and $D_{Y=0}$ are the sets of subjects with $Y=1$ and $Y=0$, respectively, n_1 and n_0 are the sizes of these sets and $I[\bullet]$ is the indicator function adjusted for ties

$$I[\hat{p}_j, \hat{p}_m] = \begin{cases} 1 & \text{if } \hat{p}_j < \hat{p}_m \\ 0.5 & \text{if } \hat{p}_j = \hat{p}_m \\ 0 & \text{otherwise.} \end{cases}$$

3.1 Proposed methodology

Assuming that the continuous predictor X is what one wishes to categorise, our proposal consists of categorising X such that the best predictive logistic model is obtained for Y. Specifically, given k the number of cut points set for categorising X in $k+1$ intervals, let us denote $\boldsymbol{v_k} = (\mathfrak{x}_1, \ldots, \mathfrak{x}_k)$ the vector of k cut points ordered from smaller to larger, and X_{cat_k} the corresponding categorised variable taking values from 0 to k. Then, what we propose is that the vector of k cut points $\boldsymbol{v_k} = (\mathfrak{x}_1, \ldots, \mathfrak{x}_k)$, which maximises the AUC of the logistic regression model shown in equation (2) is thus the vector of the optimal k cut points.

$$P(Y=1|X_{cat_k}) = logit^{-1}(\beta_0 + \sum_{q=1}^{k} \beta_q 1_{\{X_{cat_k}=q\}}). \tag{2}$$

Estimation of the model in equation (2) as well as of the associated AUC can be done as presented before for the model in equation (1). However, the problem now lies in looking for the vector of the cut points which maximises the AUC. To achieve this, we propose two alternative algorithms, respectively named *AddFor* and *Genetic*.

AddFor:

Using this algorithm, one cut point is searched for at a time. In other words, it first seeks $\mathfrak{x}_1$ (in a grid of size M of equally spaced values in the range of X), such that the AUC of the logistic regression model shown in equation (2) for $k=1$ will be maximised. Once $\mathfrak{x}_1$ has been selected, it is fixed and the algorithm proceeds to seek $\mathfrak{x}_2$ (in the grid of size M) ($\mathfrak{x}_2 \neq \mathfrak{x}_1$), so as to ensure that the AUC of the model in (2) for $k=2$ will be maximised. The process is then repeated until the vector of k cut points, $\boldsymbol{v_k} = (\mathfrak{x}[1], \ldots, \mathfrak{x}[k])$, has been obtained, with $\mathfrak{x}[o]$ denoting the o-th ordered cut point.

Genetic:

Using Genetic Algorithms, the most widely known type of evolutionary algorithm[23], this method simultaneously finds the vector of k cut points, $\boldsymbol{v_k} = (\mathfrak{x}_1, \ldots, \mathfrak{x}_k)$, which maximises the AUC of the logistic regression model in equation (2). Evolutionary algorithms are inspired by the concept of natural evolution. The underlying idea is that, given a population of individuals, environmental pressure leads to survival of the fittest, leading in turn to a rise in the overall fitness of the population. In a more mathematical context, given a function to be maximised (fitness function), a collection of heuristic rules are used to modify a population of possible solutions in such a way that each generation of potential solutions, tends to be, on average, better than its predecessor. The measure whether one potential solution is better than another is the potential solution's fitness value. In our case, the AUC is the selected fitness function to be maximised and the optimal cut points would be then the best possible solution.

For both algorithms, the methodology above has been presented (for ease of notation and illustration) for the particular case of the categorisation of a continuous covariate X in a univariate logistic regression model. Nevertheless it can be easily extended to the categorisation of a continuous covariate X in a multiple logistic regression model. Suppose that along with the predictor variable X we want to categorise, a set of other p predictors, $Z_1, \dots, Z_p$, are of interest. Then, the categorisation of X in a multivariable setting including the p predictors, will be that for which the AUC of the multiple logistic regression model in equation (3) is maximised.

$$P(Y = 1|(Z_1, \dots, Z_p, X_{cat_k})) = logit^{-1}(\beta_0 + \sum_{r=1}^{p} \beta_r Z_r + \sum_{q=p+1}^{p+k} \beta_q 1_{\{X_{cat_k}=q-p\}}). \tag{3}$$

3.2 Optimism Correction for the AUC

When implementing the algorithms presented in the previous section, the obtained AUC may be biased upward when the same data set is used to: a) fit the logistic regression model (involved in the cut point selection process); and, b) compute the AUC[24]. In our setting, the aim was to look for the vector $\boldsymbol{v_k}$ that maximises the AUC of the corresponding logistic model. Thus, the overestimation of the AUC may have an impact in the maximisation process itself and therefore on the selection of the optimal cut points. Several approaches for correcting the bias of the estimated discriminative ability of a predictive model have been proposed in the statistical literature[25,26]. In this work, the proposal is based on the bootstrap bias correction method proposed by Steyerberg[26]. Moreover, the bias correction procedure was performed at two different levels. In the first approach, the bias correction was performed during the selection of the optimal cut points. In the second approach, however, the bias correction procedure was applied once the optimal cut points had been selected. Appendix A of the web supporting material shows the results of a simulation study performed to evaluate the impact of the bias correction approaches (at first and second level) on the selection of the optimal cut points. As can be observed on the results, both approaches provide similar results. Hence, and due to computational cost savings, we propose to correct the AUC at the end of the cut point selection process. Specifically, in this approach, the bootstrap bias correction method can be described at follows:

Step 1. Categorise the predictor variable on the basis of the original sample $\{(x_i, y_i)\}_{i=1}^{N}$ and compute the corresponding AUC. Let's denote this *apparent* AUC as $\widehat{AUC}_{app}$.

Step 2. For $b = 1, \ldots, B$, generate the bootstrap resample $\{(x^*_{ib}, y^*_{ib})\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample, and categorise the bootstrapped predictor $\{x^*_{ib}\}_{i=1}^N$ on the basis of the optimal cut points obtained in Step 1.

Step 3. Fit the logistic regression model to the bootstrap resample with the categorised version of the predictor and compute the corresponding AUC, $\widehat{AUC}^b_{boot}$ for $b = 1, \ldots, B$.

Step 4. Obtain the predicted probabilities for the original sample based on the fitted logistic regression model obtained in Step 3 and compute the AUC. Let's denote this AUC as $\widehat{AUC}^b_o$ for $b = 1, \ldots, B$.

Once the above process has been completed, the optimism O of the original AUC is calculated as follows

$$O = \frac{1}{B}\sum_{b=1}^{B} |\widehat{AUC}^b_{boot} - \widehat{AUC}^b_o|$$

and the bias corrected AUC is then computed as $\widehat{AUC}_{app} - O$.

Finally, we would like to point out that in order to mimic the study design, it is advisable that the resampling procedure described in Step 2 be done according to the design of the study. For instance, for a case-control study, data should be resampled separately within cases and controls. Moreover, if the data are clustered, the resampling units should be the clusters.

3.3 Selection of the number of cut points

To determine the optimal number of cut points we studied two possible approaches. The first approach is based on the difference between the bias-corrected AUCs obtained for $k = l$ and $k = l + 1$ cut points. To determine the need for an extra optimal cut point, we propose to compute the confidence interval (CI) for this difference. Once the CI has been computed, an extra cut point is considered to be needed as long as the CI does not contain the zero. Specifically, in this paper bootstrap-based methods[27] are proposed for constructing the CIs. The procedure can be summarised as follows:

1. For $v = 1, \ldots, V$, generate the bootstrap resample $\{(x^*_{iv}, y^*_{iv})\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample.

2. Compute the bias corrected AUC for the categorised variable for $k = l$ and $k = l + 1$ and denote it as $\widehat{AUC}^*_{l,v}$ and $\widehat{AUC}^*_{l+1,v}$ respectively. The bias corrected AUC is computed as explained in Section 3.2, but using for Step 1 the optimal cut points obtained for $k = l$ and $k = l + 1$ on the basis of the original sample.

3. Compute the difference between the bias-corrected AUCs obtained for $k = l + 1$ and $k = l$

$$\widehat{AUC}^*_{Diff,v} = \widehat{AUC}^*_{l+1,v} - \widehat{AUC}^*_{l,v}.$$

Once the above process has been completed, the $(1 - \alpha)$ % limits for the CI for the difference are given by

$$\left(\widehat{AUC}^{\alpha/2}_{Diff}, \widehat{AUC}^{1-\alpha/2}_{Diff} \right)$$

where $\widehat{AUC}^{p}_{Diff}$ represents the p-percentile of the estimated $\widehat{AUC}^*_{Diff,v}$ $(v = 1, \dots, V)$.

The second criterion used to evaluate the need for an extra optimal cut point was the integrated discrimination improvement (IDI) index, proposed by Pencina et al[28] which in our setting can be defined as shown in equation (1) in Pepe et al.[29]:

$$IDI = E[p_{l+1} - p_l | Y = 1] - E[p_{l+1} - p_l | Y = 0],$$

where $p_k = P(Y = 1|X_{cat_k})$.

The IDI is a useful measure to compare and assess the improvement in terms of risk prediction of two predictive models. Accordingly, in our particular setting, the IDI can be a useful measure to evaluate the improvement offered by adding an extra cut point. In particular, we propose the criterion that an extra cut point is needed as long as an statistically significant IDI is obtained when comparing the fitted logistic regression models obtained with $k = l$ and $k = l + 1$ cut points.

4 Validation study

This section reports the results of a simulation study conducted to analyse the empirical performance of the methods described in Section 3 above. Validation was provided at two different levels, i.e., in a theoretical setting and in a backward process. Both settings are explained in detail below.

All computations were performed using the (64 bit) R 3.0.1 software package[30].

4.1 Scenarios and set-up

Theoretical validation:
In the first setting, the predictor variable X was simulated from a normal distribution separately in each of the populations defined by the outcome ($Y = 0$ and $Y = 1$), i.e., $X|(Y = 0) \simeq N(\mu_0, \sigma_0)$ and $X|(Y = 1) \simeq N(\mu_1, \sigma_1)$. It should be noted that, when σ_0 and σ_1 are equal, the linear relationship between X and the logistic function holds. Moreover it can be shown that for k cut points, the theoretical location of the optimal cut points can be obtained[12], as well as the AUC associated with the corresponding categorical covariate. Accordingly, the aims of this simulation study were

twofold:- a) to compare the cut points obtained with the proposed methodology and the theoretical optimal cut points; and b) to compare the obtained bias corrected AUC and the theoretical one. The most general results are presented in the main manuscript. Nevertheless, more specific results for different scenarios and sample sizes are presented in the web appendixes B and C. Specifically, in the simulations presented in this manuscript, we considered $X|(Y = 0) \simeq N(0, 1)$, $X|(Y = 1) \simeq N(1.5, 1)$. The simulations were done assuming the same number of individuals in $Y = 0$ and $Y = 1$ and a total sample sizes of $N = 500$ and $N = 1000$. As far as the number of cut points is concerned, $k = 1, 2$ and 3 was considered. Finally, for the *AddFor* algorithm grid sizes of $M = 100$ and $M = 1000$ were used. In all cases, $B = 50$ was considered for the AUC bias corrected procedure. Sample sizes of $N = 500$ and $N = 1000$ were selected to ensure a requirement commonly used on the specific framework of prediction models[26]. Nevertheless, the performance of the proposed methodology was also verified for smaller sample sizes. Detailed results can be seen in Appendix B of the supporting web material. $R = 500$ and $R = 1000$ replicates of simulated data were performed. Both number of replicates provided similar results (not shown). Accordingly, all results shown here are based on $R = 500$ replicates.

As pointed out before, under this setting ($\sigma_1 = \sigma_0$) the relationship between the predictor X and the *logit* transformation of the response Y is linear. Nevertheless, the performance of the proposed methodology when the relationship is not linear was also assessed by comparing both algorithms in a controlled situation. Results are shown in Appendix C of the supporting web material.

Backward validation:

In the second setting, we envisaged simulating a continuous variable X starting from a categorical variable whose cut points had been scientifically pre-established and assuming that they represent an underlying continuum variable. The aim was to test whether the cut points obtained by applying the proposed methodology to the continuous variable were similar to the original cut-points. For the purpose, we considered the data set available at the IRYSS-COPD study[19]. In this data set we selected the variable forced expiratory volume in 1 second in percentage ($FEV_{1\%}$) which is a clinical variable whose categorisation into four categories (mild ≥ 80, moderate $[50-80)$, severe $[30-50)$ and very severe < 30) is well established thanks to previous research in the field[31]. This variable was available in the data set both in the continuous and the categorical versions for a total number of $L = 2069$ patients.

To simulate the continuous covariate $FEV_{1\%}$ we propose a bootstrap method starting from the original categorical and continuous versions of $FEV_{1\%}$. Let us denote X the original continuous $FEV_{1\%}$ variable and X_{cat} the categorised variable taking values from 0 to 3, which correspond to mild, moderate, severe and very severe categories respectively. For each $l = 0, \ldots, 3$, consider d_{ls} as the s-th decile of X when $X_{cat} = l$. For each $u = 1, \ldots, U$ and $l = 0, \ldots, 3$, we generated the bootstrap sample $\{x^*_{iu}\}_{i=1}^{L_l}$ by drawing a sample of size L_l with replacement from the original sample $\{x_i\}_{i=1}^{L_l}$, where L_l denotes the number of individuals in the l-th category ($L = \sum_{l=0}^{3} L_l$). We considered d^*_{ls} as the average of the U bootstrap deciles of each category, i.e, $d^*_{ls} = \frac{1}{U}\sum_{u=1}^{U} d^u_{ls}$. The continuous variable X_{sim} was simulated assuming a uniform distribution in the interval $(d^*_{l(s-1)}, d^*_{ls})$, enclosed by the lower and upper limits of each category.

Additionally the dichotomous response variable Y was simulated according to the two scenarios shown in Table 1 trying to mimic two possible real situations. In Scenario I patients are distributed as 35%, 30%, 20% and 15% in mild, moderate, severe and very severe categories respectively. In contrast, in Scenario II, only a 3% of patients belongs to the mild category. Additionally, the percentage of individuals with $Y = 1$ (denoted as diseased), changes from Scenario I to Scenario II.

Table 1: Backward validation study: total distribution of individuals in the four categories of forced expiratory volume in 1 second in percentage ($FEV_{1\%}$) and distribution of diseased individuals in each category, under both scenarios.

	Scenario I		**Scenario II**	
$FEV_{1\%}$ **[0,100]**	**Total**	**Diseased**	**Total**	**Diseased**
Mild [80,100]	35%	5%	3%	0%
Moderate [50,80)	30%	20%	30%	4.5%
Severe [30,50)	20%	25%	47%	8.6%
Very severe [0,30)	15%	40%	20%	14.2%

For each of the scenarios, $R = 500$ replicates were conducted for total sample sizes of $N = 500$ and $N = 1000$, and $B = 50$ as in the previous setting and $U = 10.000$ bootstrap resamples were used. Optimal cut points were sought using the *Genetic* and *AddFor* algorithms, the latter with grid sizes of $M = 100$ and $M = 1000$.

4.2 Results

Theoretical validation:

Figure 1 depicts the boxplot of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithms, different sample sizes and number of cut points. As can be observed, the cut points obtained by the *Genetic* or *AddFor* algorithms were close to the theoretical optimal cut points, with both algorithms presenting a low bias. The corresponding detailed numerical results are shown in Table 2. Under this scenario, the theoretical optimal cut points are $\boldsymbol{v_1} = (0.77)$, $\boldsymbol{v_2} = (0.23, 1.27)$ and $\boldsymbol{v_3} = (-0.07, 0.75, 1.57)$ for $k = 1, 2, 3$ number of cut points respectively. Note that the average of the estimated cut points across simulated data sets was very similar for both algorithms with this values being very close to the theoretical optimal cut points.

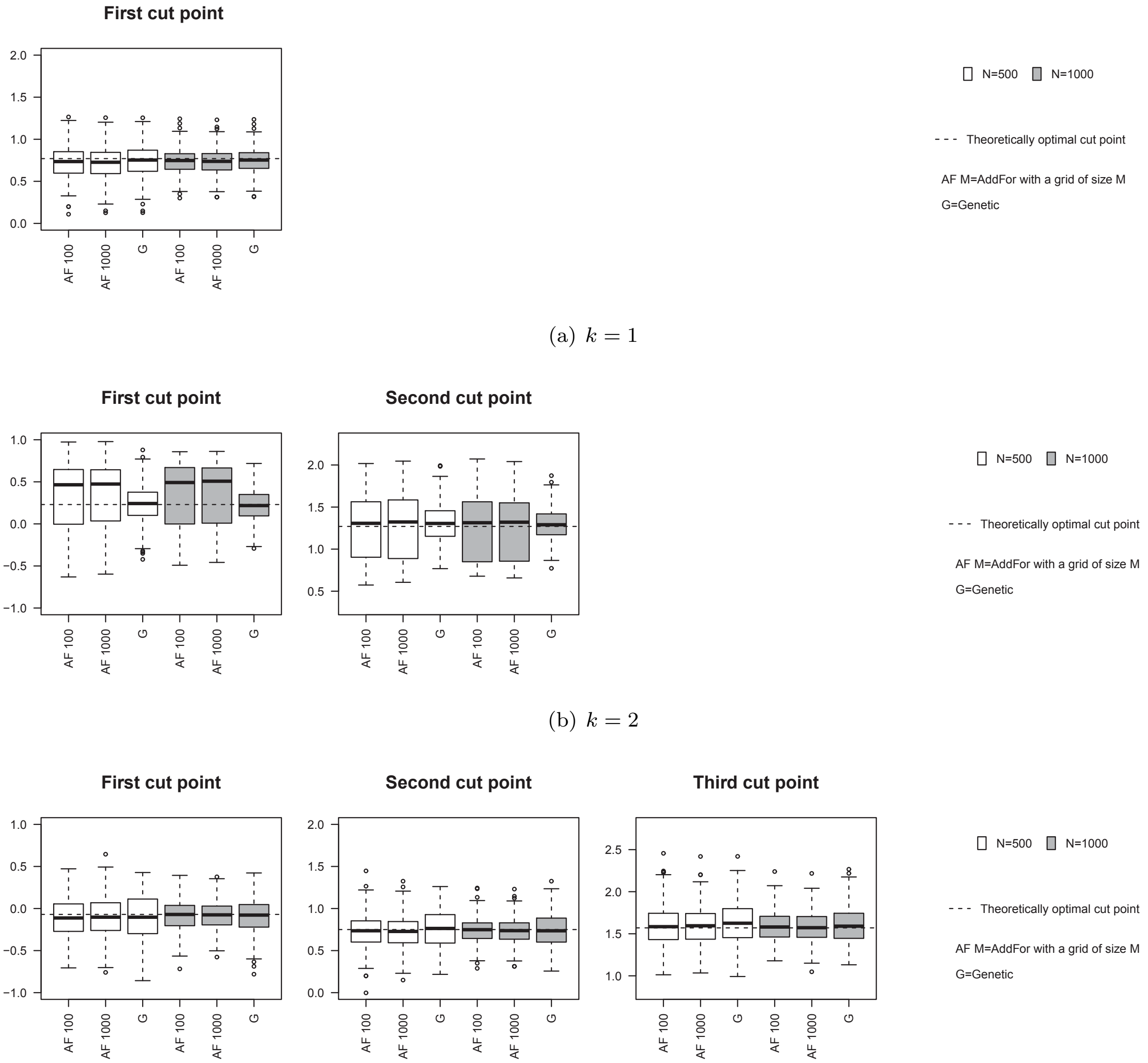
As expected, the differences with respect to the theoretical optimal cut points were smaller when the sample size increased from 500 to 1000. For example, for $k = 3$ cut points, the average of the cut points obtained with the *Genetic* method across all replicates were $\bar{\boldsymbol{v}}_{\mathbf{3}} = (-0.11, 0.76, 1.63)$ and $\bar{\boldsymbol{v}}_{\mathbf{3}} = (-0.09, 0.74, 1.61)$ for sample sizes of 500 and 1000 respectively, while with the *Addfor* algorithm and a grid of size 1000 they were $\bar{\boldsymbol{v}}_{\mathbf{3}} = (-0.11, 0.73, 1.60)$ and $\bar{\boldsymbol{v}}_{\mathbf{3}} = (-0.08, 0.74, 1.58)$. It should be noted that, when the desired number of cut points was 2, the *AddFor* did not perform as well as the *Genetic* algorithm. While the former located only one of the two optimal cut points, the latter managed to approximate both cut points. For instance, for a sample size of 500 and $k = 2$, the bias obtained for the estimated cut points were $(0.00, 0.04)$ using the *Genetic* method, and $(0.11, -0.01)$ using the *AddFor* method with a grid of size 1000.

In Table 3, the average, bias and standard deviation of the bias corrected AUC values over 500 simulated data sets is given, for each of the proposed algorithms, different sample sizes and number of cut points. Note that the AUC values obtained were almost unbiased, being the bias obtained less or equal to 0.02 when $k = 1$ was chosen. Additionally, the *Genetic* approach generally provided slightly higher AUC values than the *AddFor* algorithm. However, when the *AddFor* grid size was increased from 100 to 1000, the obtained results were almost the same as those obtained with the *Genetic* algorithm. For instance, for a sample size of 500 and $k = 3$ number of cut points, the average of bias corrected AUCs was 0.831, 0.834 and 0.835 for the *AddFor* with grid sizes of 100 and 1000 and the *Genetic* algorithm respectively.

It should be noted that, the *Genetic* algorithm is computationally more expensive than the *AddFor* algorithm. Detailed information about computational cost and convergence to the AUC of the continuous variable for larger k values is given in the Appendix D of the supporting web material.

Backward validation:

The backward-validation simulation study showed that both the *AddFor* and *Genetic* methods were able to detect the original cut points. This can be observed in Figure 2 where the boxplots of the estimated optimal cut points based on 500 simulated data sets are depicted, for each of the proposed methods and different sample sizes. The corresponding numerical results are shown in Table 4 were the average of the optimal cut points together with the original cut points are shown. Note that the cut points obtained with the *Genetic* method were slightly closer to the original cut points than the ones obtained with the *AddFor*. For instance, under Scenario I and a sample size of 1000, the average of the estimated optimal cut points obtained by the *Genetic* method were 32.03, 53.98 and 77.99 while the ones obtained with the *AddFor* method with a grid of size 1000 were 32.96, 56.91 and 77.17. It is worth remembering that the original three cut points were 30, 50 and 80. Table 4 also shows that under Scenario II, only 2 of the original three cut points were detected. The percentage of patients with a $FEV_{1\%}$ of over 80 was less than 3%, and none of them was diseased. Hence, having so few individuals with values above 80, the method was not able to detect that cut point. In this situation, the first two cut points were retained and the original cut points of 30 and 50 were detected.



(a) $k = 1$

(b) $k = 2$

(c) $k = 3$

Figure 1: Boxplot of the estimated optimal cut points based on 500 simulated data obtained according to the theoretical optimal cut point validation study and comparison with the theoretically optimal cut point ($\boldsymbol{v_1} = (0.77)$, $\boldsymbol{v_2} = (0.23, 1.27)$ and $\boldsymbol{v_3} = (-0.07, 0.75, 1.57)$). From top to bottom: (a) for $k = 1$ number of cut points; (b) for $k = 2$ number of cut points; and (c) for $k = 3$ number of cut points.

Table 2: Numerical results of the theoretical validation study. The average, bias and standard deviation of the estimated optimal cut points over 500 simulated data and $N = 500$, 1000 are shown, jointly with the theoretical cut points.

	Sample size $N = 500$									**Sample size** $N = 1000$									
k	*AddFor* $M = 100$			*AddFor* $M = 1000$			*Genetic*			*AddFor* $M = 100$			*AddFor* $M = 1000$			*Genetic*			**Theoretical cut point**
	av	bi	sd	av	bi	sd	av	bi	sd	av	bi	sd	av	bi	sd	av	bi	sd	
1	0.73	-0.04	0.18	0.72	-0.05	0.18	0.75	-0.02	0.18	0.74	-0.03	0.15	0.74	-0.03	0.14	0.75	-0.02	0.14	0.77
2	0.33	0.10	0.37	0.34	0.11	0.36	0.23	0.00	0.22	0.34	0.11	0.36	0.34	0.11	0.36	0.22	-0.01	0.18	0.23
	1.26	-0.01	0.36	1.26	-0.01	0.37	1.31	0.04	0.22	1.23	-0.04	0.36	1.23	-0.04	0.36	1.30	0.03	0.18	1.27
3	-0.12	-0.05	0.23	-0.11	-0.04	0.24	-0.11	-0.04	0.27	-0.08	-0.01	0.17	-0.08	-0.01	0.16	-0.09	-0.02	0.20	-0.07
	0.73	-0.02	0.19	0.73	-0.02	0.18	0.76	0.01	0.22	0.74	-0.01	0.15	0.74	-0.01	0.14	0.74	-0.01	0.19	0.75
	1.60	0.03	0.23	1.60	0.03	0.22	1.63	0.06	0.25	1.59	0.02	0.18	1.58	0.01	0.18	1.61	0.04	0.21	1.57

av: average; bi: bias; sd: standard deviation, k: number of cut points; M: grid size for the *AddFor* algorithm.

Table 3: Numerical results of the theoretical validation study. The average, bias and standard deviation of the estimated bias corrected AUC over 500 simulated data and $N = 500$, 1000 are shown, jointly with the theoretical AUC associated to the corresponding categorical covariate.

	Sample size $N = 500$									**Sample size** $N = 1000$									
k	*AddFor* $M = 100$			*AddFor* $M = 1000$			*Genetic*			*AddFor* $M = 100$			*AddFor* $M = 1000$			*Genetic*			**tAUC**
	av	bi	sd	av	bi	sd	av	bi	sd	av	bi	sd	av	bi	sd	av	bi	sd	
1	0.766	0.016	0.017	0.768	0.018	0.017	0.769	0.019	0.017	0.768	0.018	0.013	0.770	0.020	0.013	0.771	0.021	0.013	0.750
2	0.807	-0.013	0.017	0.810	-0.010	0.017	0.818	-0.002	0.016	0.807	-0.013	0.013	0.809	-0.011	0.013	0.819	-0.001	0.012	0.820
3	0.831	-0.004	0.016	0.834	-0.001	0.016	0.835	0.000	0.016	0.832	-0.003	0.012	0.835	0.000	0.012	0.836	0.001	0.012	0.835
Continuous predictor's theoretical AUC 0.855																			

av: average; bi: bias; sd: standard deviation; tAUC:theoretical AUC for each categorical variable; k: number of cut points; M: grid size for the *AddFor* algorithm.

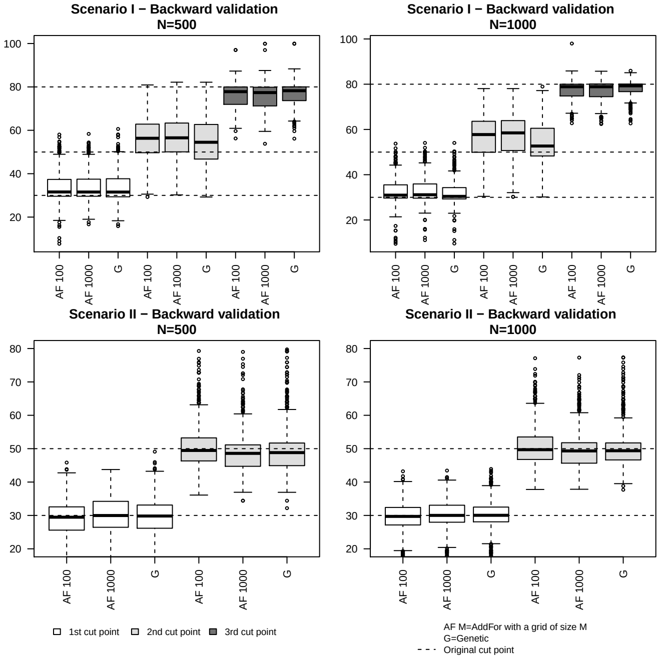


Figure 2: Boxplot of the estimated optimal cut points based on 500 simulated data sets obtained according to the backward validation study for $N = 500$ and $N = 1000$. Top row: Scenario I. Bottom row: Scenario II. Original cut points were 30, 50 and 80.

5 Software implementation

To provide the biomedical researchers with an easy-to-use tool for categorising continuous variables in prediction models, the methodology described in this paper has been implemented in the `R` programming language[30]. Specifically, an `R` package, called `CatPredi`, was created, with the *Genetic* method being implemented using the `R` package `rgenoud`[32]. The `CatPredi` package can be freely downloaded from https://sites.google.com/site/biostit/lineas-de-investigacion/software/catpredi.

By providing the dichotomous response Y, the continuous covariate which is aimed to categorise X, a set of covariates $\boldsymbol{Z}$ (if the aim is to categorise X in a multivariable setting) and k, the number of cut points, the user can choose which algorithm to use for categorising X. If a multivariable setting is chosen, the set of covariates $\boldsymbol{Z}$ can be modelled considering linear or non linear effects alternatively. In the latest, the effects are estimated using the `R` package `mgcv`[4].

The main function of the `CatPredi` package called `catpredi` returns the optimal cut points jointly with the categorised predictor variable, as well as the final model's original and bias cor-

Table 4: Results of the backward validation study: average of the estimated optimal cut points over 500 simulated data sets obtained together with the original cut points 30, 50 and 80 are shown.

	Sample size $N = 500$			**Sample size** $N = 1000$			**Original cut point**
	AddFor $M = 100$	*AddFor* $M = 1000$	*Genetic*	*AddFor* $M = 100$	*AddFor* $M = 1000$	*Genetic*	
Scenario I							
1st cut point	33.72	33.99	33.93	32.67	32.96	32.03	30
2nd cut point	55.83	56.03	56.10	56.50	56.91	53.98	50
3rd cut point	75.98	75.55	79.03	77.29	77.17	77.99	80
Scenario II							
1st cut point	29.23	30.16	29.89	29.35	30.23	30.05	30
2nd cut point	50.61	49.21	49.89	51.02	50.02	50.39	50

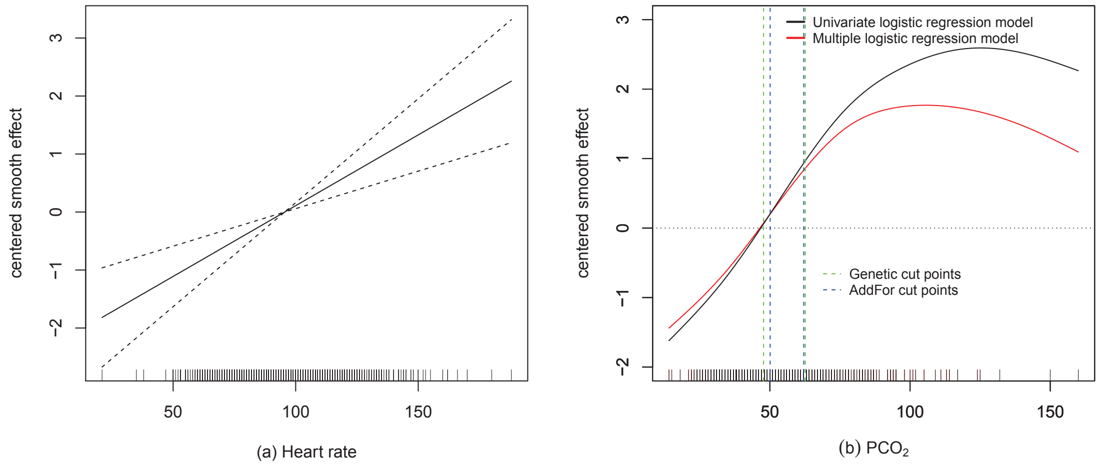
rected AUC. Additionally, it provides a graphical display of the relationship between the continuous predictor X and the response Y, estimated on the basis of a logistic GAM using the `R` package `mgcv`[4]. In this graphical display, the location of the obtained optimal cut points is also indicated. A brief detail description of the usage and arguments of this function is given below:

```
catpredi(cat.var, formula, cat.points = 1, range = NULL, data,
 method = c("addfor", "genetic"), correct.AUC = TRUE,
 control = control.catpredi())
```

- `cat.var`: name of the covariate we want to categorise.
- `formula`: this argument allows the user to specify whether the continuous predictor should be categorised in a univariable context, or in presence of other covariates or confounders, i.e in a multiple logistic regression model. For instance, `Y ∼ 1` indicates that the categorisation should be done in a univariable setting, with Y being the response variable.
- `cat.points`: number of optimal cut points to look for.
- `data`: data set to be used for the selection of the optimal cut points.
- `method`: algorithm for the selection of the optimal cut points, either *AddFor* or *Genetic*.
- `range`: range of the covariate in which to look for the cut points.
- `correct.AUC`: A logical value. If TRUE the bias corrected AUC is estimated.
- `control`: Used to set various parameters controlling the fitting process. For instance, the grid size used for the *AddFor* algorithm would be specified as `control= control.catpredi(addfor.g=1000)` for grid of size 1000.

6 Application to eCOPD data set

We applied the methodology proposed in this paper to the IRYSS-COPD study presented in Section 2. As pointed out before, preliminary analysis during the development of a prediction model for patients with eCOPD showed that clinical variables related with short term very severe evolution were the Glasgow coma scale (0: altered, 1:normal), the heart rate and the PCO_2. Moreover, this preliminary analysis also suggested that the relationship between the heart rate and the response variable short term very severe evolution appeared to be linear, while the relationship between the PCO_2 and the response variable did not. This can be seen in Figure 3 were the estimated effects of both heart rate and PCO_2 based on a logistic GAM[4] are depicted. For this reason clinical researchers decided to introduce a categorised version of the PCO_2 variable into the prediction model. However, as there was no unified criteria between clinicians about cut off points, we developed and applied the methodology presented in this work to obtain optimal cut points.



(a) Heart rate

(b) PCO_2

Figure 3: From left to right: (a) Relationship of the predictor variable heart rate with short term very severe evolution. (b) Estimated smooth relationship of the predictor variable partial pressure of carbon dioxide in the blood (PCO_2) with the response variable short term very severe evolution in a univariate logistic regression model and in a multiple logistic regression model adjusted by Glasgow and heart rate covariates, jointly with the cut points obtained with the *AddFor* and *Genetic* methods.

As a first step, we considered categorising the PCO_2 variable into 2, 3 and 4 categories in a univariable setting. To determine the optimal number of cut points we applied the two approaches presented in Section 3. For all the analyses, the implemented `catpredi` function of the `CatPredi` package was used, and both the *AddFor* and *Genetic* algorithms were applied to these data using, by way of example, the following code:

```
cat.pco2 < - catpredi(formula = poor_evolution ~ 1, cat.var="pco2",
cat.points=3, data=data.copd, method="addfor",
correct.AUC = TRUE, control = controlcatpredi(addfor.g = 1000))
```

Table 5 shows the results obtained in the categorisation of the predictor PCO$_2$ with the *AddFor* and the *Genetic* algorithms. In the case of the *AddFor* algorithm, similar results were obtained when grid sizes of 100 and 1000 were chosen, and so only the results for $M = 1000$ are reported. For each number of cut points (k =1,2 and 3), the obtained optimal cut points together with the bias corrected AUC are reported. Additionally, the difference of the bias corrected AUCs, as well as the IDI indexes when compared models with 1 and 2 and 2 and 3 cut points are shown.

Table 5: Results obtained in the categorisation of the predictor variable partial pressure of carbon dioxide in the blood of the IRYSS-COPD study in a univariable setting.

Method	k	cut points	Bias corrected AUC	AUC difference (95% CI^*)	IDI (95% CI)
Addfor $M = 1000$	1	50.1	0.674		
				0.022 (0.011 , 0.036)	0.016 (0.008,0.024)
	2	50.1; 62.08	0.696		
				0.016 (-0.002 , 0.045)	0.001 (-0.0003,0.002)
	3	45.86; 50.1; 62.08	0.712		
Genetic	1	50.87	0.674		
				0.032 (0.010 , 0.065)	0.016 (0.008,0.025)
	2	47.74; 62.64	0.706		
				0.006 (-0.002 , 0.025)	0.0002 (-0.0003,0.001)
	3	34.06; 47.52; 62.58	0.713		

* 95% bootstrap confidence interval (CI) based on the percentile method for $V = 100$ number of bootstrap resamples.
k: number of cut points; M: grid size for the *AddFor* algorithm;
IDI: integrated discrimination improvement; AUC: area under the ROC curve.

As can be observed, the cut points obtained with the *Genetic* and *AddFor* algorithms were quite similar, with those obtained for $k = 1$ being 50.10 and 50.87, those obtained for $k = 2$ being (50.10, 62.08) and (47.74, 62.64) and those obtained for $k = 3$ being (45.86, 50.10, 62.08) and (34.06, 47.52, 62.58), using the *AddFor* with a grid of size 1000 and the *Genetic* algorithms respectively. If we selected the cut points based on percentiles, the results would be: 44 for $k = 1$ (median), (40, 49) for $k = 2$ (0.333 and 0.666 percentiles) and (38, 44 , 53) for $k = 3$ (quartiles). The corresponding bias corrected AUCs for these categorical variables based on percentiles would be: 0.644 for $k = 1$, 0.676 for $k = 2$ and 0.688 for $k = 3$.

In the case of the *Genetic* algorithm, bias corrected AUCs of 0.674, 0.706 and 0.713 were obtained for $k = 1, 2$ and 3 respectively. A difference (95% bootstrap CI) of 0.032 (0.010, 0.065) was obtained between AUCs for $k = 2$ and $k = 1$ cut points and a difference of 0.006 (-0.002, 0.025) between AUCs for $k = 3$ and $k = 2$ cut points. The IDI obtained when passed from $k = 1$ to $k = 2$ cut points was of 0.016 (p-value = 0.0002). However, when passed from $k = 2$ to $k = 3$ cut points the IDI was of 0.0002 (p-value = 0.385).

In the case of the *AddFor* algorithm with a grid of size 1000, bias corrected AUCs of 0.674, 0.696 and 0.712 were obtained for $k = 1, 2$ and 3 respectively. A difference of 0.022 (0.011, 0.036) was obtained between AUCs for $k = 2$ and $k = 1$ cut points and a difference of 0.016 (-0.002, 0.045) between AUCs for $k = 3$ and $k = 2$ cut points. The IDI obtained when passed from $k = 1$ to $k = 2$ cut points was of 0.016 (p-value = 0.0002). However, when passed from $k = 2$ to $k = 3$ cut points the IDI was of 0.0007 (p-value = 0.147).

Summarising, all the results suggested that 2 was the optimal number of cut points, being the vector of optimal cut points $\hat{\boldsymbol{v}}_{\mathbf{2}} = (47.74, 62.64)$ or $\hat{\boldsymbol{v}}_{\mathbf{2}} = (50.1, 62.08)$ if the *Genetic* or *AddFor* algorithm was chosen. These can be seen in Figure 3(b). In either case, no significant differences were observed between the AUC values obtained with the categorical variable and the bias corrected AUC yielded by the original continuous predictor (using a logistic GAM) resulting in 0.707. Consequently, we obtained a categorical version of the continuous PCO_2 variable whose discriminative ability compared to the continuous original version did not decrease significantly with a 95% bootstrap CI of (-0.03,0.02) and (-0.02,0.04) with the *Genetic* and *AddFor* algorithms respectively. In addition, we obtained a 92% agreement between the categorical variables achieved with the *AddFor* and *Genetic* algorithms, measured by Cohen's weighted kappa[33], with a 95% CI of (0.91, 0.93). These results were face-validated by the clinicians involved in the IRYSS-COPD study.

Finally, we considered to categorise the predictor variable PCO_2 in a multivariable setting adjusted by the other predictor variables considered by clinicians, which were Glasgow coma scale and heart rate. The cut points obtained for $k = 2$ were (47.03, 62.08) and (47.33, 62.54) with the *AddFor* and *Genetic* algorithms respectively. In this case the effect of the other covariates in the multiple logistic regression model did not change the optimal cut points obtained for the PCO_2 covariate. This result could be explained by looking at the estimated effects shown in Figure 3(b) where it can be observed that the shape of the relationship between the continuous predictor and the response variable does not change from the univariable to the multivariable setting.

7 Discussion

The disadvantages of categorising continuous variables, such as the loss of information and statistical power, have been reported by a number of authors[34,35]. From a practical point of view, however, categorisation may be useful for ease of interpretation and application, especially when the aim is to apply the results in daily clinical practice, or even helpful in cases where outliers are present. Indeed, in routine practice continuous clinical predictors are usually available but health professional's decision-making tends to be based on patient risk classification, which can be seen as a categorisation of the original continuous predictor. Bearing this in mind, i.e., the fact that decisions are made on the basis of the risk classification of patients, we feel that a prediction model should be in line with the decision-making process and to incorporate categorical predictor variables into the prediction model may be a way of doing so.

As pointed out in the introduction, previous work has been done on the categorisation of continuous variables but most of these studies have sought to dichotomise the continuous variable and have not taken the search for more than one cut point into account. Our study indicates, however, that the loss of information in terms of discriminative ability could be very high when a single cut point is considered.

The methodology presented in this paper allows for the selection of more than one cut point. The advantages that our proposal presents with respect to previously published proposals for the selection of more than one cut point are: a) it requires no distributional assumptions and can be used in any situation regardless of the distribution of the original continuous predictor[12]; and b) it provides the objectivity afforded by an automatic method as opposed to the subjectivity of relying on a graphical display[13]. Furthermore, our approach has been developed so that a continuous predictor variable can be categorised both in a univariable or a multivariable context, depending on what the underlying setting is for each data set (univariable or multivariable) as proposed by Mazumdar et al[36]. Although in the application to the IRYSS-COPD study the cut points obtained for the PCO_2 covariate in the univariable and multivariable settings were almost the same, this need not always be so. The cut points obtained in the multivariable setting may differ to those obtained in the univariate model. For example, if the relationship between the continuous predictor and the response variable is different in a univariable or a multivariable setting, then the optimal cut points may be different. This could happen for example when confusion predictors are present in the multiple logistic regression model. Hence, in contrast to other categorisation methods, the proposed methodology thus enables a continuous variable to be categorised before or during the development of a prediction model, thereby allowing for the incorporation of potential confounders.

The simulation study shows, that under the theoretical hypothesis, our approach yields the optimal location of the cut points. Additionally, the results obtained suggest that the cut points obtained correspond to the change at risk of having the outcome of interest. Indeed and according to clinicians criteria, in the application of the method to the IRYSS-COPD study, the cut points obtained for the clinical variable PCO_2, classified patients into low, moderate and high risk of short term very severe evolution. The proposed methods thus provide a classification of patients in terms of risk, which is precisely what is desirable in clinical practice for decision-making.

Two alternative algorithms are proposed in this paper. Despite the fact that the results are similar, one must bear in mind that the *AddFor* algorithm seeks the second cut point once the first has been fixed. Consequently, the selection of the first cut point has an influence on the consecutive cut points, which at times may lead to a non-optimal selection of cut points. We think that this could be solved by improving the algorithm with a backward/forward correction, adjusting the first cut point after the second has been selected and so on, which is part of our future research. In general, as long as it is computationally feasible, we recommend the use of the *Genetic* algorithm.

Note that the proposed methodology consists on categorising a continuous predictor variable X. In practice, data may be clumped at a point or points which may lead to have ties on the data. If this happens, we should focus on sample size, percentage of ties and more concretely on the number of unique values the predictor variable takes. If these are a few, then the predictor variable would be far from being a continuous variable and hence applying our methods is not advisable in such cases. We conducted a simulation study in which we considered different levels of digits preference for a continuous normally distributed variable in the same conditions of the theoretical validation study and a sample size of $N = 500$ (data not shown). The results suggested that whenever the number of unique values of the predictor variable are over 60 we can completely recommend the application of the proposed methodology.

The main limitation of the proposal is that it does not include a methodology which obtains the optimal number of cut points. The researcher must select in advance the number of cut points, or compare the performance of the categorised variable for different number of cut points. We have seen that the predictive ability will increase as the number of cut points increases but also, that it converges to the theoretical predictive ability of the continuous predictor. We are aware that in theory the optimal number of cut points for the categorisation of a continuous variable does not exists, since above all the possible number of cut points the best option would be the continuous variable. However, in clinical practice categorical versions of the continuous variables are usually preferred without having always clear which is the best number of categories to be used. It is necessary to find a balance between the clinical sense of the categories and the minimal loss of information. Therefore, we have proposed two approaches to select the best number of cut points, which we have applied to the IRYSS-COPD data. In our opinion, future work on the development and validation of these naive approaches to select the optimal number of cut points is called for.

On the other hand, the proposed approach is based on the fit of a logistic regression model from the data available. Bearing this in mind, if there are missing data on the predictor variable we aim to categorise, these would not be taken into account in the categorisation process, since the logistic model would be fitted only on the complete observations. If missing data are completely at random, its effect will be determined by the reduction in the available sample size. The simulation study suggests that the method proposed performs satisfactorily regardless of the sample size. However, when missing data are not at random, it is known that it has an impact on the modelling process itself. Consequently, it will also have an impact on the selection of the optimal cut points.

Another limitation is that only one unique continuous covariate can be categorised at a time. We are working on the extension of these methods to categorise two predictor variables at a time, considering the influence they have on each other and with the response variable. In the meantime, if a researcher is interested on the categorisation of two continuous variables, we recommend to proceed as follows. Consider categorising each variable at a time in a multiple logistic regression model in which the second variable is modelled using non linear effects. In this way, optimal cut points for each of the continuous variables would be obtained but adjusted by the effect of the second variable. Once cut points for both variables have been obtained, categorise and combine them in a final model. Although this is not as categorising both variables at a time, it can be seen as an easy to apply first approximation.

To summarise, we propose a method for categorising continuous predictors in a logistic regression model which provides the optimal location of the cut points and two algorithms for its implementation. The two algorithms have been compared and validated. The aim of this methodology is to provide optimal cut points but taking into account that they must be always validated with clinical researchers. Hence, when this method was applied to a real data set, the resulting cut points were face-validated by clinicians. Furthermore, the proposed categorical predictors were seen to perform as successfully as the continuous variables. Finally, an `R` package called `CatPredi` has been implemented which leads to an easy use of this methodology in practice.

Acknowledgements We would like to thank the referees for providing thoughtful comments and suggestions which led to substantial improvement of this manuscript.
Funding This study was supported by grants MTM2010-14913, 2012111008, IT620-13, UFI11/52, MTM2011-28285-C02-01 and MTM2013-40941-P. The work of IB was supported by grant GIU10/21 from the University of the Basque Country UPV/EHU. MXRA thanks the Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273). The collection of the COPD data used for this study was supported in part by grants from the Fondo de Investigación Sanitaria (PI 061010, PI061017, PI06714, PI060326 and PI060664), Basque Country Regional Health Authority, Galdakao Hospital Research Committee and the thematic networks-Red IRYSS (Investigación en Resultados y Servicios Sanitarios)- of the Instituto de Salud Carlos III (G03/220).
Conflict of interest *The authors declare that there are no conflicts of interest.*

References

1. Altman DG and Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* 1998; 52: 289-303.

2. Royston P, Altman DG and Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006; 25 :127-141.

3. Hastie T and Tibshirani R. *Generalized additive models.* London: Chapman & Hall, 1990.

4. Wood SN. *Generalized additive models: an introduction with R.* London: Chapman & Hall, 2006.

5. Turner E, Dobson J and Pocock J. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiol Perspect Innov* 2010; 7: 9.

6. Bennette C and Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol* 2012; 12: 21.

7. Lim BL and Kelly AM. A meta-analysis on the utility of peripheral venous blood gas analyses in exacerbations of chronic obstructive pulmonary disease in the emergency department. *Eur J Emerg Med* 2010; 17: 246-248.

8. Mazumdar M and Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med* 2000; 19: 113-132.

9. Lausen B and Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Comput Stat Data Anal* 1996; 21: 307-326.

10. Hin LY, Lau TK, Rogers MS, et al. Dichotomization of continuous measurements using generalized additive modelling - application in predicting intrapartum caesarean delivery. *Stat Med* 1999; 18: 1101-1110.

11. Magder LS and Fix AD. Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. *J Clin Epidemiol* 2003; 56: 956-962.

12. Tsuruta H and Bax L. Polychotomization of continuous variables in regression models based on the overall C index. *BMC Med Inform Decis Mak* 2006; 6: 41.

13. Barrio I, Arostegui I, Quintana JM, et al. Use of generalised additive models to categorise continuous variables in clinical prediction. *BMC Med Res Methodol* 2013; 13: 83.

14. Harrell FE, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA J Am Med Assoc* 1982; 247: 2543-2546.

15. Buist AS, Vollmer WM and McBurnie MA. Worldwide burden of COPD in high-and low-income countries. Part I. The Burden of Obstructive Lung Disease (BOLD) Initiative. *Int J Tuberc Lung Dis* 2008; 12: 703-708.

16. Esteban C, Quintana JM, Moraza J, et al. Impact of hospitalisations for exacerbations of COPD on health-related quality of life. *Respir Med* 2009; 103: 1201-1208.

17. Rabe KF, Hurd S, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 2007; 176: 532-555.

18. Pauwels RA and Rabe, KF. Burden and clinical features of chronic obstructive pulmonary disease (COPD). *Lancet* 2004; 364: 613-620.

19. Quintana JM, Esteban C, Barrio I, et al. The IRYSS-COPD appropriateness study: objectives, methodology, and description of the prospective cohort. *BMC Health Serv Res* 2011; 11: 322.

20. Teasdale G and Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974; 304: 81-84.

21. McCullagh P and Nelder JA. *Generalized linear models.* London: Chapman & Hall, 1989.

22. Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press, 2003.

23. Eiben AE and Smith JE. *Introduction to evolutionary computing.* Springer, 2003.

24. Copas JB and Corbett P. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 2002; 89: 315-331.

25. Airola A, Pahikkala T, Waegeman W, et al. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput Stat Data Anal* 2011; 55: 1828-1844.

26. Steyerberg EW. *Clinical prediction models. A practical approach to development, validation, and updating.* New York: Springer, 2009.

27. Efron B and Tibshirani RJ. *An introduction to the bootstrap.* New York: Chapman & Hall, 1993.

28. Pencina MJ, D'Agostino RB and Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008; 27: 157-172.

29. Pepe MS, Feng Z and Gu JW. Comments on Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyondby MJ Pencina et al., Statistics in Medicine (DOI: 10.1002/sim. 2929). *Stat Med* 2008; 27: 173-181.

30. R Core Team. R: A Language and Environment for Statistical Computing. http://www.R-project.org/ (2013).

31. Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease . *http://www.goldcopd.com/* updated 2013.

32. Mebane WR and Sekhon JS. Genetic optimization using derivatives: the rgenoud package for R . *J Stat Softw* 2011; 42: 1-26.

33. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213.

34. Taylor JMG and Yu M. Bias and efficiency loss due to categorizing an explanatory variable. *J Multivar Anal* 2002; 83: 248-263.

35. Altman DG. Categorizing Continuous Variables. In: *Encyclopedia of Biostatistics*. John Wiley & Sons, 2005.

36. Mazumdar M, Smith A and Bacik J. Methods for categorizing a prognostic variable in a multivariable setting. *Stat Med* 2003; 22: 559-571.