Selecting the number of categories of the lymph node ratio in cancer research: a bootstrap-based hypothesis test

Journal Title

XX(X):2–41

©The Author(s) 0000

Reprints and permission:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/ToBeAssigned

www.sagepub.com/

SAGE

Irantzu Barrio ‡ [1,4]

Javier Roca-Pardiñas [2]

Inmaculada Arostegui[1,3,4]

[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] **BCAM- Basque Center for Applied Mathematics**

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

‡Corresponding author: E-mail: irantzu.barrio@ehu.eus, Tel.: +34-946012504, Address: Departamento de Matemática Aplicada, Estadística e Investigación Operativa. Facultad de Ciencia y Tecnología. Universidad del País Vasco UPV/EHU. Barrio Sarriena s/n. 48940 Leioa.

Prepared using sagej.cls

Abstract

The high impact of the lymph node ratio as a prognostic factor is widely established in colorectal cancer, being used as a categorized predictor variable in several studies. However, the cut-off points as well as the number of categories considered differ considerably in the literature. Motivated by the need to obtain the best categorization of the lymph node ratio as a predictor of mortality in colorectal cancer patients, we propose a method to select the best number of categories for a continuous variable in a logistic regression framework. Thus, to this end, we propose a bootstrap-based hypothesis test, together with a new estimation algorithm for the optimal location of the cut-off points called *BackAddFor*, which is an updated version of the previously proposed *AddFor* algorithm. The performance of the hypothesis test was evaluated by means of a simulation study, under different scenarios, yielding type I errors close to the nominal errors and good power values whenever a meaningful difference in terms of prediction ability existed. Finally, the methodology proposed was applied to the CCR-CARESS study where the lymph node ratio was included as a predictor of five-year mortality, resulting in the selection of three categories.

Keywords: categorization; prediction models; cut-off point; bootstrap

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

Colorectal cancer is an important cause of mortality all over the world, with more than 800 000 cases in 2015 [1] and it is among the most commonly occurring cancers in men and women [2]. Prediction models to estimate the risk of mortality or relapse of the disease need to be developed to enhance prevention and early detection. These models identify the tumour-, genetic-, and patient-associated risk factors, which allow to groups of patients to be categorized depending on high or low risk levels.

Several risk prediction models for mortality and other adverse events, such as recurrence or complications, have been developed [3,4], and the impact of different prognostic factors has been discussed [5–7]. In this context, the high prognostic impact of the lymph node ratio (LNR), i.e. the ratio of metastatic to examined lymph nodes, is widely established in colorectal cancer [8–11]. In all these studies, the LNR was categorized; however, the cut-off points as well as the number of categories considered differed among the studies. Berger et al. [8] divided the LNR into four categories using the quartiles as cut-off points. On the other hand, De Ridder et al. [9] and Rosenberg et al. [10] used statistical methods to select the optimal cut-off points. While the former considered a dichotomized version of the LNR, choosing a cut-off point that maximized the Nagelkerke's r^2 index, the later used the minimum p-value approach of the log-rank test to group the LNR into four categories. Therefore, not only did the methods to select the cut-off points and their location differ among the studies, but also the number of categories to be considered.

On the other hand, the classification and staging of cancer enables the physicians to stratify patients, which leads to better treatment decisions and the development of a

common cancer treatment strategies[12]. Therefore, an efficient classification of lymphatic involvement is crucial to define the prognosis of the disease and define an adecuate theraphy for patients depending on the stage they belong to. Moreover, the most clinically useful cancer staging system is the tumour node metastasis (TNM) system created by the American Joint Committee on Cancer (AJCC) and the International Union for Cancer Control (UICC)[12]. The TNM system classifies cancers by the size and extend of the primary tumour (T), involvement of regional lymph node (N) and the presence or absence of distant metastasis (M). The classification is based on categorizations and combinations of the three components in a specific way for each anatomic site, allowing the stratification of patients with similar prognosis.

The Results and Health Services Research in Colorectal Cancer (CCR-CARESS) project is a prospective cohort study of new cases of colorectal cancer patients with five years of follow-up[13]. One of the main purposes of the CCR-CARESS study was the development of clinical prediction models and scores for mortality or tumour recurrence. In this study, categorization of some of the continuous predictors was performed during the modelling phase for reasons such as lack of linearity, model interpretability as well as ease of the punctuation of the scores derived from the models or compliance with clinical practice criteria[7]. Therefore, the selection of a method for categorization, the location of the cut-off points, and the optimal number of categories were relevant issues that arose while the clinical prediction models were being developed in the CCR-CARESS study. However, for variables such as LNR, which is an important variable in the staging system, a wide variety of either number and location of cut-off points have been used in

the literature and therefore researchers wanted to study the best possible categorization for this variable.

A methodology to select the optimal cut-off points to categorize a continuous predictor in the context of logistic regression has been previously proposed. It was based on the maximal discrimination ability of the model measured by the area under the receiver operative characteristic (ROC) curve (AUC)[14]. Furthermore, this methodology has also been extended for use in the Cox proportional hazards regression model[15]. Although this methodology allows selecting any possible number of cut-off points, either in a univariate or a multiple context, in many circumstances the number of categories in which to categorize the predictor variable is unclear. Moreover, the disadvantages of categorizing a continuous variable, such as, for instance, the loss of information or statistical power, have been widely discussed[16]. However, if both the number and location of the cut-off points are appropriately searched for, it is possible to minimize the loss in predictive ability with respect to the continuous predictor modelled by means of a spline function. Along these lines, and motivated by the need to find the most appropriate categorization of the LNR in the CCR-CARESS study, we were interested in a methodology that provides not only the optimal location of a pre-selected number of cut-off points, but also the best number of cut-off points to look for. Therefore, the aim of the current study is to propose a bootstrap-based hypothesis test to select the best number of categories in which a continuous predictor variable should be categorized in the context of a logistic prediction model.

The rest of the paper is organized as follows. Section 2 provides a description of the CCR-CARESS study of patients with colorectal cancer which motivated the development of the methodology presented in this paper. Section 3 outlines some notation and background along with a new algorithm to estimate the optimal cut-off points. In Section 4, a bootstrap-based hypothesis test is proposed for selecting the best number of categories. In Section 5, we present the results from two simulation studies that were conducted to assess the performance of the methodologies presented in Sections 3 and 4, respectively. Section 6 describes the application of the proposed methodology to the CCR-CARESS study data set. Finally, the paper closes with a discussion in Section 7 in which the main findings are reviewed and conclusions are drawn.

Motivating data: The CCR-CARESS study

CCR-CARESS is a prospective multi-center cohort study of patients diagnosed with colorectal cancer who had undergone surgical interventions between June 2010 and December 2012 and consented to participate in the study. Subjects were followed up for up to five years. The 22 participating hospitals represent six regions in Spain, and they operate under the Spanish National Health Service. Patients' selection criteria, explicit definition of diagnosis, and patient recruitment were explained in detail in the study protocol[13].

Variables collected include pre-intervention background, sociodemographic parameters, hospital admission records, biological and clinical parameters, treatment information, and outcomes up to five years after surgical intervention. LNR, defined as the ratio

of tumour-infiltrated lymph nodes to total number of resected lymph nodes, was collected after the surgical intervention.

The role of LNR as a predictive factor in the prognosis for colorectal cancer patients has been widely discussed previously[5,8–10], especially for colon cancer. Therefore, LNR was one of the predictive variables that was considered when developing a clinical prediction model of mortality in the CCR-CARESS study. Previous results derived from the study included different categorizations and transformations of the LNR in mortality prediction of colon cancer[7,17]. The absence of consensus related to the best categorization in terms of both the location of the cut-off points and the number of categories, and the controversy around this issue, was the motivation for this work. In the current study, only patients with a diagnosis of colon cancer were selected from the CCR-CARESS study. Further, patients with incomplete follow-up as regards to mortality were excluded (161, 8.1%) for this study as the necessary information was not available to be able to include them in the 5-year predictive model. A brief description of the main variables of the CCR-CARESS study is reported in the supplementary material (Web Appendix A).

Models and estimation algorithms

Let Y be a binary (0/1) response variable and $(\boldsymbol{Z}, X)$ a vector of associated covariates with $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$ and X a continuous variable. In this context, the generalized linear model (GLM) with a logistic link is commonly used to predict the probability of success ($Y = 1$) considering the values taken by the covariates. Let us use the notation $P(\mathbf{Z}, X) = P(Y = 1|\mathbf{Z}, X)$. Traditionally, the effect of the continuous covariate X on

the logit$(P(\mathbf{Z}, X))$ is considered to be linear. However, that effect may be nonlinear but can be approximated by a piecewise constant relationship. In that case, the effect of X would be given by the following expression:

$$\text{logit}(P(\mathbf{Z}, X)) = \log \frac{P(\mathbf{Z}, X)}{1 - P(\mathbf{Z}, X)} = \delta_0 + \sum_{r=1}^{q} \delta_r Z_r + \sum_{s=1}^{k} \beta_s \left(1_{c_s < X \leq c_{s+1}}\right) \quad (1)$$

where $(\delta_0, \delta_1, \ldots, \delta_q; \beta_1, \ldots, \beta_k)$ are unknown coefficients associated with the effect of the covariates, and $\boldsymbol{c} = (c_1, \ldots, c_k)$ is a vector of k cut-off points (ordered from lowest to highest) which define the $k + 1$ intervals for the variable X (considering $[min(X), c_1]$ as the reference category). Moreover, for simplicity of notation, we consider that $c_{k+1} = \infty$.

Given a sample $\{(\boldsymbol{z}_i, x_i, y_i)\}_{i=1}^{n}$ and fixing the vector of cut-off points $\boldsymbol{c}$, from the expression in equation (1), the estimation of the true probability $p(\mathbf{z}, x)$ is given by

$$\hat{p}(\boldsymbol{c}, \mathbf{z}, x) = \frac{\exp\left(\hat{\delta}_0(\boldsymbol{c}) + \sum_{r=1}^{q} \hat{\delta}_r(\boldsymbol{c}) z_r + \sum_{s=1}^{k} \hat{\beta}_s(\boldsymbol{c}) \left(1_{c_s < X \leq c_{s+1}}\right)\right)}{1 + \exp\left(\hat{\delta}_0(\boldsymbol{c}) + \sum_{r=1}^{q} \hat{\delta}_r(\boldsymbol{c}) z_r + \sum_{s=1}^{k} \hat{\beta}_s(\boldsymbol{c}) \left(1_{c_s < X \leq c_{s+1}}\right)\right)} \quad (2)$$

where $\widehat{\delta}_0(\boldsymbol{c}), \widehat{\delta}_1(\boldsymbol{c}), \ldots, \widehat{\delta}_q(\boldsymbol{c})$ and $\widehat{\beta}_1(\boldsymbol{c}), \ldots, \widehat{\beta}_k(\boldsymbol{c})$ are the estimated coefficients obtained by maximum likelihood (using the iterative weighted least squares local scoring algorithm[18]).

The discriminatory ability of a logistic model is commonly measured by the AUC[19,20]. Once the estimated probabilities $\hat{p}(\boldsymbol{c}, \mathbf{z}_i, x_i)$ for $i = 1, \ldots, n$, have been computed,

the estimation of the true AUC ($\widehat{AUC}(\boldsymbol{c})$) can be obtained using the Mann-Whitney statistic[21] as follows:

$$\widehat{AUC}(\boldsymbol{c}) = \frac{1}{n_0 n_1} \sum_{j \in D_{Y=0}} \sum_{m \in D_{Y=1}} I[\hat{p}(\boldsymbol{c}, \mathbf{z}_j, x_j), \hat{p}(\boldsymbol{c}, \mathbf{z}_m, x_m)], \tag{3}$$

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 $I[\bullet] = I[p(\boldsymbol{c}, \mathbf{z}_j, x_j), p(\boldsymbol{c}, \mathbf{z}_m, x_m)]$ is the indicator function adjusted for ties:

$$I[\bullet] = \begin{cases} 1 & \text{if } p(\boldsymbol{c}, \mathbf{z}_j, x_j) < p(\boldsymbol{c}, \mathbf{z}_m, x_m) \\ 0.5 & \text{if } p(\boldsymbol{c}, \mathbf{z}_j, x_j) = p(\boldsymbol{c}, \mathbf{z}_m, x_m) \quad \forall j \in D_{Y=0} \text{ and } \forall m \in D_{Y=1} \\ 0 & \text{otherwise.} \end{cases}$$

Note that the obtained $\widehat{AUC}(\boldsymbol{c})$ depends on $\boldsymbol{c}$. In practice, the locations of the cut-off points are unknown, and thus it is necessary to estimate them. Barrio et al.[14] proposed and compared two algorithms named *AddFor* and *Genetic* to estimate the optimal locations of the cut-off points $\boldsymbol{c} = (c_1, \ldots, c_k)$, considering the maximal AUC.

In the *AddFor* algorithm, one cut-off point is searched for at a time. To be specific, the steps of the proposed algorithm are as follows: first, considering $k = 1$, the first cut-off point is obtained as the one that maximizes the AUC of the probabilities given in equation (2) for $k = 1$, i.e. $\hat{c}_1 = \text{argmax}_c \widehat{AUC}(c)$; fixing $\hat{c}_1$, the second cut-off point is obtained (for $k = 2$) as $\hat{c}_2 = \text{argmax}_c \widehat{AUC}(\hat{c}_1, c)$; fixing $(\hat{c}_1, \hat{c}_2)$, the third cut-off point is obtained as $\hat{c}_3 = \text{argmax}_c \widehat{AUC}(\hat{c}_1, \hat{c}_2, c)$; and the process is then repeated until the

vector of k cut-off points $\hat{\boldsymbol{c}} = (\hat{c}_1, ..., \hat{c}_k)$ has been obtained for a previously fixed value of k.

On the other hand, the *Genetic* algorithm simultaneously finds the vector of k cut-off points by using an evolutionary algorithm[22] which looks for the vector $\boldsymbol{c} = (c_1, \ldots, c_k)$ that maximizes the fitness function $\widehat{AUC}(\boldsymbol{c})$.

As discussed in Barrio et al., the *AddFor* algorithm may lead to a non-optimal vector of cut-off points because the selection of each cut-off point is influenced by the preceding selected ones[14]. The authors recommend the use of the *Genetic* algorithm as far as it is computationally achievable. However, for large data sets or when the process needs to be incorporated into a more complex procedure, especially if it comprises a bootstrap estimation method, this may not be feasible. Therefore, in this paper, we propose a new estimation algorithm called *BackAddFor*, which is an updated procedure of the previously described *AddFor* algorithm. The steps of this new algorithm are as follows:

Initialize: Compute the initial estimates $(\hat{c}_{0,1}, ..., \hat{c}_{0,k})$.

Step 1: Cycle $j = 1, \ldots, k$ calculating the update

$$\hat{c}_j = \text{argmax}_c \widehat{AUC}(\hat{c}_1, \ldots, \hat{c}_{j-1}, c, \hat{c}_{0,j+1}, \ldots, \hat{c}_{0,k})$$

Step 2: Repeat **Step 1** replacing $(\hat{c}_{0,1}, ..., \hat{c}_{0,k})$ by $(\hat{c}_1, ..., \hat{c}_k)$ until the difference $\widehat{AUC}(\hat{c}_1, ..., \hat{c}_k) - \widehat{AUC}(\hat{c}_{0,1}, ..., \hat{c}_{0,k})$ is zero.

In Step 1, different strategies can be considered for the initialization of the cut-off points, such as: (a) random selection of k cut-off points in the range of X; (b) a grid of size k of equally spaced values in the range of X, and (c) quantiles of X. In our experience

we have seen no difference between any of the three options, in practice the algorithm converges in few iterations. Nevertheless, we propose to use a random selection prior to the selection of a k-size grid, to avoid possible problems with local minimums.

In addition, as detailed in Step 2, the initial cut-off points are updated after searching for the final cut-off points. The difference between the *BackAddfor* and the *Addfor* algorithm is that the latest does not update the cut-off points assuming that the optimal cut-off point when $k = 1$ is also one of the optimal cut-off points when $k > 1$. As seen in Barrio et al.[14] this does not always have to be the case, and therefore we expect that by allowing an update of the cut-off points the results will be improved.

Model Selection

In the previous section, we proposed a new algorithm to look for the vector of the cut-off points $\boldsymbol{c} = (c_1, \ldots, c_k)$ given that the number of categories, $k + 1$, was previously fixed. In this section, we propose a procedure that will help to determine the optimum number of categories to be considered in the categorization of a continuous predictor variable.

To this end, given k, the number of cut-off points, consideration will be given to a test for the null hypothesis:

$$H_0(k) : \text{logit}(P(\mathbf{Z}, X)) = \delta_0 + \sum_{r=1}^{q} \delta_r Z_r + \sum_{s=1}^{k} \beta_s \left(1_{c_s^0 < X \leq c_{s+1}^0}\right) \tag{4}$$

for some k cut-off points $\boldsymbol{c}^0 = (c_1^0 < \ldots < c_k^0)$ (considering $c_{k+1}^0 = \infty$) versus the alternative hypothesis

$$H_1(k) : \text{logit}(P(\mathbf{Z}, X)) = \delta_0 + \sum_{r=1}^{q} \delta_r Z_r + \sum_{s=1}^{k+1} \beta_s \left(1_{c_s^1 < X \leq c_{s+1}^1}\right) \tag{5}$$

for some $k+1$ cut-off points $\boldsymbol{c}^1 = (c_1^1 < \ldots < c_{k+1}^1)$ (considering $c_{k+2}^1 = \infty$).

To test H_0, we propose the use of a statistic based on the increment of the predictive loss function, which is defined as

$$T = \widehat{AUC}(\hat{c}_1^1, ..., \hat{c}_{k+1}^1) - \widehat{AUC}(\hat{c}_1^0, ..., \hat{c}_k^0), \tag{6}$$

where $\widehat{AUC}(\hat{c}_1^0, ..., \hat{c}_k^0)$ and $\widehat{AUC}(\hat{c}_1^1, ..., \hat{c}_{k+1}^1)$ are the estimated AUCs for the models under the null and the alternative hypothesis, respectively. It must be remarked that if T takes a high enough value, the test would decide in favour of the alternative hypothesis, while if T is small, the null hypothesis would not be rejected. Thus, the test rule for checking $H = H_0$ with significance level α is that the null hypothesis is rejected if T is larger than its upper α-percentile. It is well known that in these kind of tests it is difficult to analytically find the exact null distribution of the test statistic, and resampling methods such as the bootstrap introduced by Efron[23] can be applied instead. To be specific, to obtain these critical values, we apply the following binary bootstrap procedure[24,25]:

Step 1. Estimate the null and alternative models based on the original sample.

Step 2.1 Obtain the null estimated cut-off points $\hat{\boldsymbol{c}}^0 = (\hat{c}_1^0, \ldots, \hat{c}_k^0)$ and the associated null estimated probabilities $\hat{p}_i^0 = \hat{p}(\hat{\boldsymbol{c}}^0, \mathbf{z}_i, x_i) \quad \text{for} \quad i = 1, \ldots, n.$

Step 2.2 Obtain the estimated cut-off points $\hat{\boldsymbol{c}}^1 = (\hat{c}_1^1, \ldots, \hat{c}_{k+1}^1)$ and the associated estimated probabilities $\hat{p}_i^1 = \hat{p}(\hat{\boldsymbol{c}}^1, \mathbf{z}_i, x_i) \quad \text{for} \quad i = 1, \ldots, n.$

Step 3. Obtain the value of the T statistic using the original sample.

Step 4. For $b = 1, \ldots, B$, generate bootstrap samples $\left\{\mathbf{x}_i, y_i^{\bullet b}\right\}_{i=1}^{n}$ with $y_i^{\bullet b} \sim Bernoulli(\hat{p}_i^0)$, and compute the bootstrap statistics $T^{\bullet b}$:

$$T^{\bullet b} = \widehat{AUC}(\hat{c}_1^{1,\bullet b}, \ldots, \hat{c}_{k+1}^{1,\bullet b}) - \widehat{AUC}(\hat{c}_1^{0,\bullet b}, \ldots, \hat{c}_k^{0,\bullet b})$$

where $(\hat{c}_1^{0,\bullet b}, \ldots, \hat{c}_k^{0,\bullet b})$ and $(\hat{c}_1^{1,\bullet b}, \ldots, \hat{c}_{k+1}^{1,\bullet b})$ are the estimated cut-off points under H_0 and H_1, respectively, obtained with the bootstrap sample.

Step 5. Finally, the decision rule for each T consists of rejecting the null hypothesis if $T > \hat{T}^{1-\alpha}$, where $\hat{T}^{1-\alpha}$ is the empirical $(1-\alpha)$-percentile of the values $T^{\bullet 1}, \ldots, T^{\bullet B}$ obtained before.

We propose this procedure as a useful tool to select the most suitable number of cut-off points. If $H_0(k)$ is not rejected, our recommendation is that only k cut-off points should be considered. Otherwise, the test is repeated with $k+1$ cut-off points until the null hypothesis is not rejected. For example, if $H_0(1)$ is not rejected, just one cut-off point is recommended for inclusion in the model. If this hypothesis is rejected, it will be required to test $H_0(2)$. If this new hypothesis is again rejected, $H_0(3)$ should be tested, and so on, until a certain $H_0(k)$ is not rejected.

Simulation study

In this section, we present several simulation studies with two different goals. On the one hand, a simulation study was conducted to compare the performance of the proposed

BackAddFor algorithm with respect to the *Genetic* and *AddFor* algorithms proposed in Barrio et al[14]. On the other hand, we carried out a simulation study to examine the behaviour of the proposed bootstrap hypothesis test presented in Model Selection Section above, under different scenarios. Both simulation studies are explained in detail below, and the results obtained are reported.

The simulation studies were performed in (64 bit) `R` 3.4.3[26]. The first simulation study was run on a workstation equipped with 24 GB of RAM, Intel Xeon E5620 processor (2.40 GHz), and the Windows 7 operating system, whereas the second simulation study was executed on a workstation equipped with 16 GB of RAM, an Intel Core i7-7700 processor (3.6 GHz), and the Windows 10 operating system. The code in `R` was observed to be slow, and therefore we wrote a much faster version of the code in `Fortran`, to obtain the results presented in the Section *Performance evaluation of the bootstrap hypothesis test* (it is possible to call it from `R`). The computational times were calculated in a computer equipped with 16 GB of RAM, Intel Core i5-8400 processor (2.8 Hz), and the Windows 10 operating system.

Comparison of the estimation algorithms

Scenarios and setup:

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 $Y = 1$), under the same conditions as the ones considered in Barrio et al.[14], i.e. $X|(Y = 0) \simeq N(\mu_0 = 0, \sigma_0 = 1)$, $X|(Y = 1) \simeq N(\mu_1 = 1.5, \sigma_1 = 1)$. Note that for $\sigma_0 = \sigma_1$, the

linear relationship between X and the logistic function holds. In addition, based on the parametric method proposed by Tsuruta and Bax[27], given k cut-off points, the theoretical locations of the optimal cut-off points can be obtained, as well as the AUC associated with the corresponding categorical covariate. To be specific, in the simulations presented in this paper, we assumed the same number of individuals in $Y = 0$ and $Y = 1$. Furthermore, different sample sizes ($n = 500$ and $n = 1000$) and number of cut-off points ($k = 2$ and $k = 3$) have been considered.

The corresponding theoretical cut-off points in these scenarios were $\boldsymbol{c} = (-0.068, 0.750, 1.568)$ and $\boldsymbol{c} = (0.227, 1.274)$ with the corresponding theoretical AUC values of 0.835 and 0.820, for $k = 3$ and $k = 2$, respectively. For the *AddFor* and *BackAddFor* algorithms, a grid size of $M = 50$ was used. All results are based on $R = 500$ replicates.

The performance of each of the algorithms was evaluated by means of the bias and mean square error (MSE) of the estimated optimal cut-off points for each iteration as follows:

$$MSE_r = \frac{1}{k}\sum_{d=1}^{k}(\hat{c}_d^r - c_d)^2, \qquad r = 1, \ldots, R,$$

where $\hat{c}_d^r$ is the estimated d^{th} optimal cut-off point in the simulation r, and c_d is the d^{th} theoretical cut-off point.

Finally, the corresponding AUC for the optimal estimated vector of cut-off points with each of the three algorithms was estimated ($\widehat{AUC}(\hat{c})$). As reported in Barrio et al.[14], this AUC is overestimated because the same data is used a) to estimate $\hat{p}(\boldsymbol{c}, x)$ for each $\boldsymbol{c}$,

and b) to estimate the corresponding $\widehat{AUC}(\boldsymbol{c})$. Therefore, the AUC bias was corrected by means of the bootstrap bias-corrected method proposed by the same authors.

Results:

The numerical results obtained for the estimated optimal cut-off points obtained with the *AddFor*, *BackAddFor*, and *Genetic* estimation algorithms for different sample sizes ($n = 500$ and $n = 1000$), number of cut-off points ($k = 2$ and $k = 3$), and $R = 500$ replicates are summarized in Table 1. As can be observed, for $k = 2$, the *BackAddFor* algorithm exhibits good performance, similar to *Genetic*, and notably improves the results obtained by the *AddFor* algorithm. More specifically, the MSEs obtained with the *AddFor* algorithm are 0.136 and 0.14 for sample sizes of $n = 500$ and $n = 1000$, respectively, while those obtained with *BackAddFor* are 0.051 and 0.039, respectively, in the same scenarios. With respect to the bias, very similar results are obtained with the *BackAddFor* and *Genetic* algorithms. In particular, for $n = 500$, the bias of the first and second cut-off points are -0.022 and 0.018 with *BackAddFor* against -0.028 and 0.009 with the *Genetic* algorithm. On the other hand, for $k = 3$, very similar results in terms of bias and MSE are obtained with all the algorithms considered.

The corresponding numerical results obtained for the estimated and bias-corrected AUCs derived from each algorithm in each scenario are given in Table 2. In general, the highest AUC values correspond to the cut-off points obtained with the *Genetic* algorithm, while the lowest correspond to those derived from the *AddFor* algorithm. Although the AUCs based on *BackAddFor* are not as high as the ones obtained with *Genetic*, as mentioned above, the cut-off points estimated with both algorithms are equally accurate

in terms of bias and MSE, and improve those obtained with the *AddFor* algorithm (Table 1). In terms of the bias-corrected AUC, the results obtained are in line with those shown in Barrio et al. [14], leading to a non-bias-estimated AUC.

With regard to computation times, the *BackAddFor* requires slightly greater computation times than the *AddFor*. For example, for $k = 2$ and $n = 1000$, the *AddFor* takes on average 0.90 seconds while the *BackAddFor* takes 2.29 seconds. However, in the same scenario, when we move to the *Genetic* algorithm the computational cost is of another magnitude, requiring on average 93.95 seconds. The specific details regarding the computational times are shown in Web Appendix D.

Table 1. Numerical results of the comparison between the *AddFor*, *BackAddFor*, and *Genetic* estimation algorithms.

k	Sample Size	Algorithm	$\boldsymbol{c}$	Mean (sd)	Median	Bias	MSE (sd)
$k=2$	n=500	*AddFor*	0.227	0.282 (0.367)	0.346	0.054	0.136 (0.083)
			1.274	1.196 (0.359)	1.127	-0.078	
		BackAddFor	0.227	0.206 (0.226)	0.195	-0.022	0.051 (0.057)
			1.274	1.292 (0.226)	1.291	0.018	
		Genetic	0.227	0.199 (0.225)	0.199	-0.028	0.049 (0.056)
			1.274	1.283 (0.214)	1.270	0.009	
$k=2$	n=1000	*AddFor*	0.227	0.378 (0.363)	0.562	0.150	0.140 (0.066)
			1.274	1.265 (0.355)	1.402	-0.009	
		BackAddFor	0.227	0.209 (0.203)	0.227	-0.018	0.039 (0.051)
			1.274	1.297 (0.191)	1.316	0.023	
		Genetic	0.227	0.219 (0.188)	0.234	-0.008	0.035 (0.042)
			1.274	1.311 (0.183)	1.314	0.037	
$k=3$	n=500	*AddFor*	-0.068	-0.103 (0.219)	-0.100	-0.035	0.047 (0.046)
			0.750	0.744 (0.186)	0.747	-0.006	
			1.568	1.600 (0.238)	1.592	0.032	
		BackAddFor	-0.068	-0.107 (0.244)	-0.098	-0.039	0.062 (0.060)
			0.750	0.777 (0.220)	0.777	0.027	
			1.568	1.634 (0.267)	1.637	0.066	
		Genetic	-0.068	-0.109 (0.238)	-0.095	-0.041	0.059 (0.056)
			0.750	0.762 (0.221)	0.751	0.012	
			1.568	1.623 (0.257)	1.605	0.054	
$k=3$	n=1000	*AddFor*	-0.068	-0.112 (0.179)	-0.112	-0.044	0.029 (0.025)
			0.750	0.728 (0.138)	0.728	-0.021	
			1.568	1.579 (0.181)	1.554	0.011	
		BackAddFor	-0.068	-0.117 (0.198)	-0.115	-0.049	0.041 (0.043)
			0.750	0.748 (0.188)	0.745	0.002	
			1.568	1.610 (0.207)	1.598	0.041	
		Genetic	-0.068	-0.110 (0.192)	-0.110	-0.042	0.04 (0.039)
			0.750	0.748 (0.188)	0.738	-0.002	
			1.568	1.610 (0.209)	1.588	0.042	

A grid of size $M = 50$ was used for the *AddFor* and *BackAddFor* algorithms. k represents the number of cut-off points chosen and $\boldsymbol{c}$ the theoretical vector of the cut-off points for each value of k.

Table 2. Numerical results of the estimated and bias-corrected AUC obtained when the optimal cut-off points were obtained with the *AddFor*, *BackAddFor* and *Genetic* estimation algorithms, respectively.

k	Sample Size	Algorithm	AUC	AUC Estimation					
				Mean (sd)	95%IR	Bias	Corrected Mean (sd)	Corrected 95% IR	Corrected Bias
$k=2$	n=500	*AddFor*	0.820	0.820 (0.017)	0.783 ; 0.849	0.000	0.806 (0.018)	0.767 ; 0.837	-0.014
		BackAddFor	0.820	0.828 (0.017)	0.793 ; 0.858	0.008	0.814 (0.017)	0.777 ; 0.845	-0.005
		Genetic	0.820	0.832 (0.016)	0.798 ; 0.862	0.013	0.819 (0.017)	0.783 ; 0.850	0.000
	n=1000	*AddFor*	0.820	0.814 (0.013)	0.789 ; 0.838	-0.006	0.804 (0.013)	0.778 ; 0.829	-0.016
		BackAddFor	0.820	0.824 (0.012)	0.799 ; 0.847	0.004	0.814 (0.012)	0.789 ; 0.839	-0.005
		Genetic	0.820	0.827 (0.012)	0.803 ; 0.850	0.008	0.818 (0.012)	0.793 ; 0.841	-0.002
$k=3$	n=500	*AddFor*	0.835	0.843 (0.016)	0.810 ; 0.870	0.008	0.830 (0.017)	0.794 ; 0.859	-0.006
		BackAddFor	0.835	0.844 (0.016)	0.810 ; 0.872	0.009	0.831 (0.017)	0.795 ; 0.861	-0.004
		Genetic	0.835	0.849 (0.016)	0.816 ; 0.875	0.014	0.836 (0.017)	0.801 ; 0.864	0.001
	n=1000	*AddFor*	0.835	0.839 (0.012)	0.813 ; 0.860	0.004	0.830 (0.012)	0.804 ; 0.852	-0.006
		BackAddFor	0.835	0.840 (0.012)	0.816 ; 0.861	0.005	0.831 (0.012)	0.806 ; 0.853	-0.005
		Genetic	0.835	0.844 (0.012)	0.819 ; 0.865	0.009	0.835 (0.012)	0.810 ; 0.857	-0.001

k represents the number of cut-off points chosen and $\boldsymbol{c}$ the theoretical cut-off points for each value of k. The 95 % IR corresponds to the 2.5 and 97.5 percentile intervals.

Performance evaluation of the bootstrap hypothesis test

Scenarios and setup:

In this simulation study, the continuous covariate X was drawn from a uniform distribution $U[-2, 2]$. Different scenarios were considered to study the behaviour of the hypothesis test under different circumstances. In particular, four different settings were simulated, each with a different goal. In the first scenario (S1), fixing the number of true cut-off points as $k = 3$ (i.e. four categories), we studied the power of the hypothesis test considering different probabilities for the fourth category, which vary according to the value of a constant a. In the second scenario (S2), we considered a model with k true cut-off points to study the performance of the test as a function of k. Finally, we considered two scenarios in which the true number of cut-off points for X did not exist; in particular, we considered a linear and a quadratic relationship between the continuous covariate X and $logit(P(Y = 1|X))$ in the third (S3) and fourth (S4) scenarios, respectively. For ease of reading, scenarios S3 and S4 are shown in detail in Web Appendix B . In all cases, $R = 1000$ replicates $\{(x_i^r, y_i^r)\}_{i=1}^n$ $(r = 1, \ldots, R)$ were generated according to the corresponding model. In all the scenarios, we considered the null hypothesis $H_0(k)$ versus the alternative $H_1(k)$, using the test statistic T explained in Model Selection Section. For determining the critical values of T, we applied a bootstrap resampling approach with $B = 300$ bootstrap samples. Type I errors and power rates were calculated as the rejection proportion of $H_0(k)$ (in $R = 1000$ replicates). The test size and power were determined for different nominal levels ($\alpha = 1\%$, $\alpha = 5\%$, $\alpha = 10\%$, $\alpha = 15\%$, and $\alpha = 20\%$) and sample sizes ($n = 500$ and $n = 1000$).

S1. In this first setting, the response variable was generated according to the following model:

$$\log \frac{p(Y=1|X)}{1-p(Y=1|X)} = \begin{cases} \text{-3} & \text{if} \quad X \leq -1.25 \\ \text{-0.75} & \text{if} \quad -1.25 < X \leq -0.25 \\ 2.5 & \text{if} \quad -0.25 < X \leq 0.75 \\ 2.5 - a & \text{if} \quad X > 0.75\,. \end{cases}$$

To study the power of the test, different values of a were considered, ranging from 0.25 to 3. In particular, the values for the constant a were limited to $(0.25, 0.5, 0.75, 1, 2, 3)$, each of them leading to a different modelling distribution (see supplementary material for additional detail) and three true cut-off points $\boldsymbol{c} = (-1.25, -0.25, 0.75)$. We also considered the case for $a = 0$, where there were two true cut-off points $\boldsymbol{c} = (-1.25, -0.25)$, in order to study the type I error of the test.

S2. In the second setting, we considered k true cut-off points according to the model

$$p(Y=1|X) = \begin{cases} p_1 & \text{if} \quad X \leq c_1 \\ p_2 & \text{if} \quad c_1 < X \leq c_2 \\ \vdots & \\ p_k & \text{if} \quad c_{k-1} < X \leq c_k \\ p_{k+1} & \text{if} \quad X > c_k \end{cases}$$

where

$$\begin{cases} c_j = -2 + j\frac{4}{k+1} \\ p_j = p_{min} + (j-1)\frac{p_{max}-p_{min}}{k} \end{cases} \quad \text{for} \quad j = 1, \ldots, k+1 \quad \text{and} \quad k \leq 10$$

with $p_{min} = 0.5 - 0.05k$ and $p_{max} = 0.5 + 0.05k$. Note that the values c_j (for $j = 1, \ldots, k$) were defined to obtain an equidistant sequence of values in the interval $[-2, 2]$, leaving aside the extreme values. Similarly, the probabilities for the first and last categories were defined in such a way that the theoretical AUC for the model defined is greater as the number of cut-off points k increases.

Results:

Given the large number of settings studied, we begin by summarizing the main findings and continue by analyzing in detail the results obtained in each of the proposed scenarios. First, the results suggest that whenever there are true optimal number of cut-off points, the test performs well in general, providing type I errors close to the nominal errors (α). In addition, for a small number of cut-off points, the test has good power. However, it tends to be conservative as the number of cut-off points increases, especially if there are no differences in terms of the estimated AUCs for $k+1$ and k cut-off points. In the following, we will show in detail the results obtained in each of the scenarios studied.

S1 We first studied the type I error of the test for $a = 0$ (i.e, $k = 2$, three categories). The results obtained (expressed in %) are shown in Table 3, from which it can be seen that the test performed well in general, with type I errors proving to be relatively close to the nominal errors.

Table 3. Estimated type I error (in $\%$) for different sample sizes ($n = 500$ and $n = 1000$) and nominal levels (1%, 5%, 10%, 15%, and 20%).

	Nominal Level (α)				
Sample Size	**1%**	**5%**	**10%**	**15%**	**20%**
$n = 500$	0.6	4.9	10.8	15.8	21.4
$n = 1000$	0.8	5.6	10.7	15.7	22.2

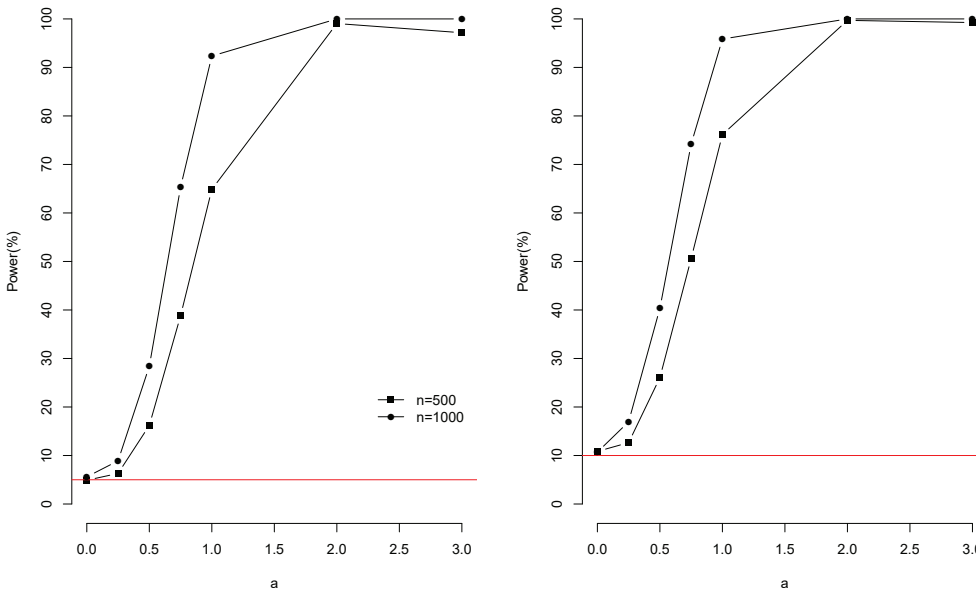

Figure 1. Percentage of rejections of the null hypothesis (power of the hypothesis test) for $\alpha = 0.05$ (left plot) and $\alpha = 0.10$ (right plot) significance levels (red line).

Figure 1 depicts the power curves for different sample sizes ($n = 500$ and $n = 1000$), values of a ($a \in \{0, 0.25, 0.5, 0.75, 1, 2, 3\}$), and significance levels ($\alpha = 0.05$ and $\alpha = 0.1$, left- and right- hand-side plots, respectively). For either sample size and significance level, when $a = 0$, the probability of rejection is approximately at the nominal level, whereas this probability rises to 1 as the value

of a increases. Although all the power curves exhibited the expected behaviour pattern, it can be observed that the power of the test strongly depends on the value of a. Thus, while the test power was very poor whenever the estimated probabilities for two categories were very similar (i.e., for small values of a), it nevertheless registered an important improvement as the value of a increased. Finally, as expected, in general, the method performed better for a sample size of $n = 1000$ than for $n = 500$.

S2 In this simulation study, we analyzed the performance of the hypothesis test as a function of the number of cut-off points k. Table 4 shows the type I errors obtained for different sample sizes ($n = 500$ and $n = 1000$), nominal levels (1%, 5%, 10%, 15%, and 20%), and number of true cut-off points k ($k = 1, \ldots, 6$) when we performed the contrast $H_0(k)$ versus the alternative $H_1(k)$. The results show that type I errors are close to the nominal levels. Nevertheless, for small sample sizes, as the number of true cut-off points increases, type I errors tend to be smaller than the nominal levels.

On the other hand, we studied the power of the test for different possible combinations of the true number of cut-off points k. Due to the large number of possible combinations, we show the results only for the case of $k = 5$ true number of cut-off points. Table 5 shows the estimated power rates when we compared the null hypothesis ($H_0(k_0)$) against the alternative ($H_1(k_0)$), for $k_0 = 1, \ldots, 4$ and data simulated based on $k = 5$ true number of cut-off points. In addition, the differences between the estimated and bias-corrected estimated AUCs for each k_0

Table 4. Estimated type I error (in %) for different sample sizes ($n = 500$ and $n = 1000$) and nominal levels (1%, 5%, 10%, 15%, and 20%) when we considered the null hypothesis $H_0(k)$ versus the alternative $H_1(k)$ for $k = 1, \ldots, 6$ under the conditions of S2.

n	**True** k	H_0 k	H_1 $k+1$	**Nominal Level** α **1%**	**5%**	**10%**	**15%**	**20%**
500	1	1	2	0.9	5.7	9.6	14.8	20.0
	2	2	3	0.8	4.4	8.7	14.0	19.3
	3	3	4	0.4	4.1	7.5	11.5	16.3
	4	4	5	0.4	3.3	7.1	11.5	16.0
	5	5	6	0.5	3.5	7.6	12.6	16.3
	6	6	7	0.5	4.6	8.5	12.1	17.4
1000	1	1	2	1.5	6.1	10.2	15.3	19.8
	2	2	3	0.5	4.2	9.4	15.3	20.1
	3	3	4	0.5	4.6	8.7	14.7	20.0
	4	4	5	0.3	4.5	10.4	15.5	20.2
	5	5	6	0.4	3.8	8.0	13.0	17.2
	6	6	7	0.6	5.6	9.9	14.7	20.2

are also reported in Table 5. The results show that even though the true number of cut-off points is 5, there are no differences in terms of the estimated AUCs, and therefore, good power rates are obtained only when 1 vs. 2 numbers of cut-off points are compared (where the differences in terms of the AUC are greater than 0.01). To study whether the low power is due to the increase in the number of cut-off points or to the need for them (in terms of the AUC), a similar simulation study has been carried out where distant probabilities have been assigned to the adjacent categories (for the details, see Web Appendix C). In this case, the power of the test is very high. Simulations have been performed for other values of k, and similar results have been obtained (results not shown). Moreover, the estimated AUCs for each k_0 have been compared with that obtained when data simulated based on $k = 5$ true number of cut-off points was modelled with a generalized

additive model (GAM), which turned out to be 0.696 (bias corrected: 0.694). As can be seen in Figure S2(c) in the supplementary material, the increment in the AUC is very slow when $k_0 \geq 3$, with similar estimated AUCs to that obtained with a GAM.

Table 5. Estimated power rates (in %) for different sample sizes ($n = 500$ and $n = 1000$) and nominal levels (1%, 5%, 10%, 15%, and 20%) when we considered $k = 5$ theoretical cut-off points under the conditions described in S2. For each comparison, the estimated AUCs and bootstrap bias-corrected AUCs are reported, together with the differences between the estimated AUCs and bias-corrected AUCs.

n	H_0 k_0	H_1 k_0+1	Nominal Level (α) **1%**	**5%**	**10%**	**15%**	**20%**	$\widehat{AUC}(k_0)$	$\widehat{AUC}^C(k_0)$	Dif	Dif^C
500	1	2	27.7	57.5	72.8	77.5	82.0	0.659	0.643	0.031	0.030
	2	3	0.5	4.3	11.6	17.6	23.7	0.690	0.672	0.011	0.010
	3	4	0.3	3.5	9.2	13.4	18.6	0.701	0.683	0.007	0.007
	4	5	0.6	3.5	9.2	12.5	17.5	0.708	0.690	0.007	0.006
1000	1	2	72.4	92.4	96.6	98.3	98.8	0.655	0.643	0.029	0.028
	2	3	4.6	21.9	39.5	48.2	55.1	0.684	0.672	0.010	0.010
	3	4	0.5	2.9	8.4	12.9	18.4	0.694	0.681	0.005	0.005
	4	5	0.4	5.1	10.7	13.2	17.4	0.699	0.686	0.004	0.004

$\widehat{AUC}(k_0)$ is the estimated AUC for k_0 number of cut-off points; $\widehat{AUC}^C(k_0)$ the bootstrap bias-corrected AUC; $Dif = \widehat{AUC}(k_0+1) - \widehat{AUC}(k_0)$ and $Dif^C = \widehat{AUC}^C(k_0+1) - \widehat{AUC}^C(k_0)$.

Application to the CCR-CARESS study

We applied the methodology proposed in this paper to the CCR-CARESS data presented in Section 2. As pointed out before, we were interested in the selection of the best number of cut-off points to categorize the predictor variable LNR. We considered the response variable *5-year mortality after surgery* and categorized the predictor variable LNR considering a univariate logistic regression model. We used the *BackAddFor* estimation algorithm with a grid of size 50. The bootstrap hypothesis test was carried out considering 400 bootstrap samples. An additional cut-off point (i.e., category) was considered statistically significant at $\alpha = 0.05$.

Table 6 shows the results obtained when the predictor variable LNR was categorized into 2, 3, 4, and 5 categories. The results suggested that the best number of cut-off points was 2 (p-value < 0.001 when 1 vs. 2 cut-off points are compared, and p-value $= 0.12$ when 2 vs. 3 cut-off points are compared). The estimated optimal cut-off points thus would be 0.06 and 0.22.

Table 6. Results obtained in the categorization of the predictor variable lymph node ratio.

Number of cut-off points (k)	AUC	AUC^C	**Estimated cut-off points** c_1	c_2	c_3	c_4	**p-value** c_k **vs.** c_{k+1}
1	0.638	0.627	0.06				< 0.001
2	0.651	0.639	0.06	0.22			0.12
3	0.653	0.643	0.04	0.10	0.22		0.95
4	0.655	0.649	0.04	0.10	0.22	0.41	

$\widehat{AUC}^C(k)$ is the bootstrap bias-corrected AUC for k cut-off points.

As can be seen in Table 6, p-values for comparing $k = 2$ vs $k = 3$ and $k = 3$ vs $k = 4$ are all greater than the previously fixed 5% significance level. However, it is not clear if

the AUCs are significantly different when $k = 2$ number of cut-off points is compared to $k = 4$ or at an even larger increment in the number of cut-off points and whether in that case different results would be obtained. In order to shed light on this issue, we have carried out all the hypothesis tests in such a way that we have contrasted $H_0 : k = k_0$ vs $H_1 : k = k_1$, where $k_0 = 1, 2, 3$ and $k_1 = k_0 + 1, \ldots, 4$. The results obtained show that one cut-off point is not enough (p-values < 0.01) and yet when contrasting $H_0 : k = 2$ cut-off points against $H_1 : k = k_1,$ for $k_1 = 3, 4$ all the p-values obtained are > 0.05, thus the optimal number of cut-off points will still be two.

Figure 2 depicts the relationship of the categorized LNR variable (2(a) two categories; 2(b) three categories; 2(c) four categories, and 2(d) five categories) together with the relationship obtained when the LNR variable was modelled considering a GAM. As can be observed, when the number of cut-off points considered is greater than 1, the estimated results obtained by the categorized variable are in line with those obtained with the GAM model. In fact, similar results are obtained in terms of discriminatory ability with the best categorization (i.e, the three categories option) and the GAM model, the estimated AUCs being 0.651 (bias corrected: 0.639) for the first and 0.654 (bias corrected: 0.651) for the GAM model, respectively (see Table 6). Nevertheless, as expected, the AUC obtained with the categorized variable is a little smaller, especially when comparing bias-corrected AUCs.

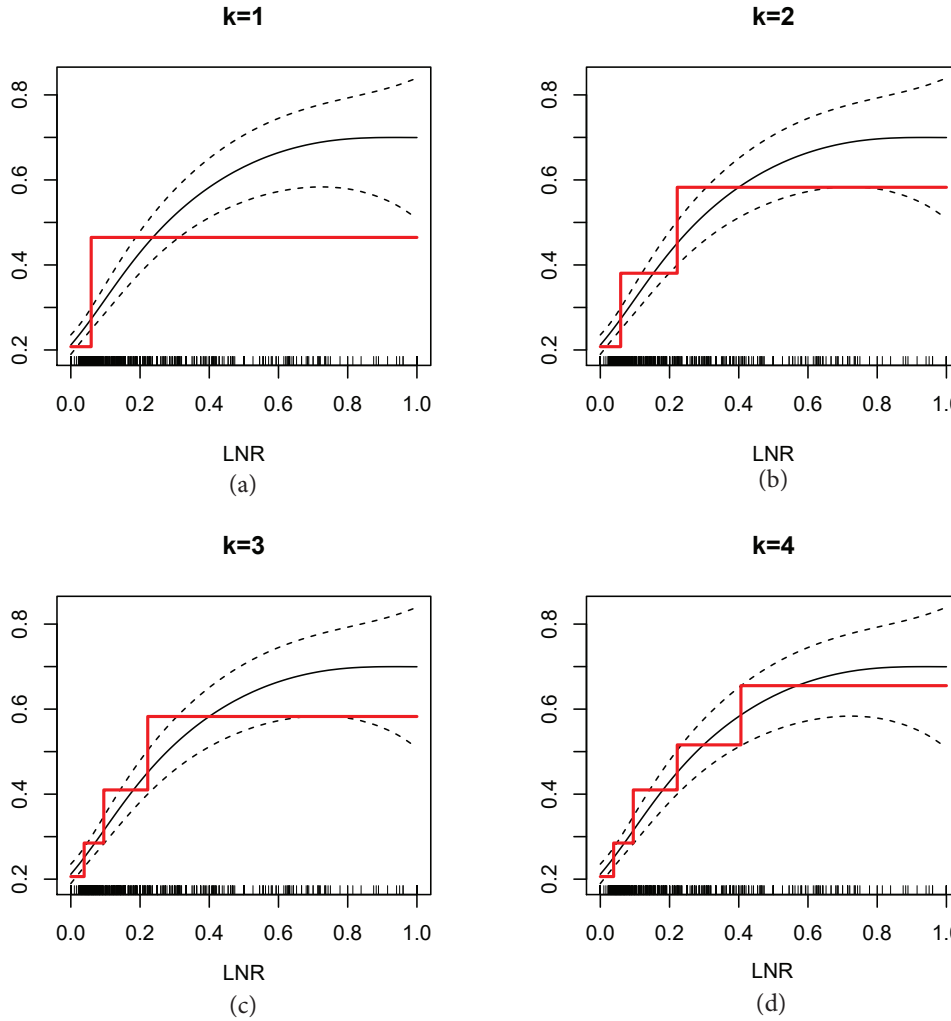

Figure 2. Graphical representation of the categorized LNR variable together with the relationship obtained with a generalized additive model (GAM).

Discussion

Although different methods have been proposed to select the optimal location of the cut-off points to categorize a continuous predictor variable[14,15,27–29], to the best of our knowledge, up to now, no approaches have been proposed in the literature to select the number of categories. Nevertheless, in the context of regression splines, the optimal allocation and selection of the number of knots has been adequately discussed in the literature[30–32]. Different algorithms have been proposed to select the optimal number of knots by minimizing (or maximizing) a certain statistical measure. For instance, Valenzuela et al.[32] proposed an algorithm to simultaneously optimize the placement and number of knots in smoothing splines by using a multi-objective genetic algorithm. Returning to the context of categorization, the need for further research into how to determine the best number of categories has already been mentioned. Barrio et al.[14] considered two procedures to select the best number of cut-off points for the application to a real data set. However, as they stated in the discussion, further work was needed on that. Therefore, in this work, we have proposed a bootstrap-based hypothesis test to compare k against $k+1$ number of categories (which also allows to compare k vs $> k$ number of cut-off points). We have conducted a simulation study with a variety of scenarios in which we have studied different relationships between the continuous covariate and the response variable. Although the scenarios are very different from one another, similar results have been obtained overall. On the one hand, type I errors very close to the nominal errors have been obtained. On the other hand, good power rates (or rejection percentages in the Linear and Quadratic scenarios) have been obtained

regardless of the value of k, provided that the differences in the AUC have been greater than 0.01.

In addition, we have proposed a new estimation algorithm called *BackAddFor*, which, according to the results obtained in our simulation studies, improves the performance of the *AddFor* algorithm whenever the performance of the latter is not accurate, and has a similar performance as the *Genetic* algorithm. Therefore, given that *BackAddFor* is computationally more efficient than *Genetic*, the former is the algorithm we have considered to develop the bootstrap hypothesis test. It is worth mentioning that unlike the proposals for the selection of the optimal number of knots in splines, we have not considered optimizing jointly the number and the location of the cut-off points. We see three reasons for this. On the one hand, in practice, there are situations in which the number of cut-off points is previously known so that joint optimization in this case would not be necessary; on the other hand, given that $AUC(k)$ is an increasing function but with a growth that tends to zero as k increases, we consider that a hypothesis contrast is needed to decide at which value of k that increase in AUC ceases to be statistically significant; and finally, we found it to be a very computationally expensive approach. However, a comparison of the differences between using the bootstrap hypothesis test and a joint optimization approach is of great interest as future work. At this point it is worth mentioning that we have selected the AUC (numerically equivalent to the c-index [33,34] in the logistic regression framework) as a measure of the discrimination ability of the categorized variable because it is the most commonly used parameter in practice. Although other parameters have been also proposed in the literature such as the effect

size and the overlap coefficient[35], these were not suitable to measure the discrimination ability of the categorized variable in this context.

Further, a bootstrap resampling procedure has been used to implement the test that will help determine the number of categories. In particular, the test statistic we have proposed consists of the increment obtained for the estimated AUC when the number of categories increases by one. We have seen that the differences between the estimated AUCs ($\widehat{AUC}(k+1) - \widehat{AUC}(k)$) and the bias-corrected AUCs ($\widehat{AUC}^{C}(k+1) - \widehat{AUC}^{C}(k)$) are approximately the same (simulation results not shown); therefore, we have not considered the need to approach the test from the standpoint of the difference of the corrected AUC, because we believe that it only increases the computational cost.

The methodology presented in this paper was applied to the CCR-CARESS study data set, and the LNR continuous variable was categorized to predict 5-year mortality in a logistic regression model. The results suggested that the best number of categories was three, with 0.06 and 0.22 being the optimal cut-off points. In addition, the bias-corrected estimated AUC obtained for this categorization proposal was 0.639. The cut-off points as well as the number of categories obtained were different from those previously used in the literature, where there was no prior consensus. In fact, Rosenberg et al.[10] identified three cut-off points (0.17, 0.41, and 0.69) which in our data set lead to an AUC of 0.612. In addition, Berger et al.[8] categorized the LNR based on quantiles, which differed from the LNR quantiles in our data set. Thus, because the distribution of the LNR variable may differ from one study to another, the optimal cut-off points may differ too. Therefore, we

propose that the optimal cut-off points and number of categories should be checked with an appropriate methodology depending on the settings of the study.

Finally, note that the proposed procedure for determining the number of categories consists of testing multiple hypotheses, where a set of k p-values corresponding to k null hypotheses, $H_0(1), \ldots, H_0(k)$, are given. To be specific, the most appropriate number of cut-off points k is obtained using the following algorithm. Step 1: Initialize with $k = 1$; Step 2: test $H_0(k)$; if the null hypothesis is rejected, then set $k = k + 1$ and repeat Step 2; otherwise k is considered as the best number of cut-off points. Therefore, the main limitation of this study is that the hypothesis test proposed may imply a multiple testing procedure. Nevertheless, unlike what we might expect, we have observed that the type 1 error does not increase when we perform the whole procedure, and that the method does not seem to select more cut-off points than necessary. Thus no cut-off points are detected that do not exist (results not shown). Even so, we consider it interesting and necessary for future work to study theoretically or empirically the type 1 error and the power of the whole process.

In the present work, the categorization of a continuous variable has been studied in a univariate model for simplicity when defining the simulation study. However, it is directly applicable to the context where there are multiple covariates in the model. However, in the case in which it is desired to categorize more than one variable at the same time, although the methodology is theoretically applicable, it is computationally unfeasible. We are currently working on alternatives to make simultaneous categorization viable. Therefore, in the case where it is required to categorize and select the number of optimal

categories for more than one variable, we propose to apply the method to one at a time, but including a smooth effect of the other covariate in the model.

In summary, we have proposed a methodology that allows one to select the number of categories whenever a predictor variable is to be categorized in a logistic regression setting. According to the simulation studies considered, the hypothesis test proposed has type I errors close to the nominal values and good power rates whenever the differences in terms of the AUC between two adjacent number of categories is larger than 0.01. In addition, the results obtained are consistent in the various situations analyzed. Nevertheless, although we have considered different scenarios representing a variety of relationships between the covariate and the response variable, we are aware that we have not studied every possible situation. Finally, a categorization of the LNR in a logistic regression model for 5-year mortality has been provided to clinical researchers with a minimum loss of discriminatory ability when compared to a GAM.

Funding:

This research was supported by the Basque Government through the Consolidated Research Group MATHMODE (IT1294-19) from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco, the BERC 2018-2021 programme and the SPRI Elkartek project 3KIA (KK-2020/00049) ; by the Spanish Government through the Ministerio de Ciencia, Innovación y Universidades: BCAM Severo Ochoa accreditation SEV-2017-0718 and by Ministerio de Economía y Competitividad and FEDER under research grants MTM2014-55966-P, MTM2016-74931-P and MTM2017-89422-P; and by Xunta de Galicia GRC ED431C 2016/040 and 2016-2019 ED431G/02.

Financial support for data collection was provided in part by grants from the Instituto de Salud Carlos III, (PS09/00314, PS09/00910, PS09/00746, PS09/00805, PI09/90460, PI09/90490, PI09/90453, PI09/90441, PI09/90397, and the thematic networks REDISSEC - Red de Investigación en Servicios de Salud en Enfermedades Crónicas), co-funded by European Regional Development Fund/European Social Fund (ERDF/ESF "Investing in your future"); and the Research Committee of the Hospital Galdakao.

Acknowledgments

We gratefully acknowledge the clinical researchers from the 22 participating hospitals for giving us the opportunity to participate in the data analysis of the CCR-CARESS study that motivated this work and also all the patients who participated in the study. In addition, we would like to thank Editage (www.editage.com) for English language editing.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Allen C et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study. *JAMA Oncology* 2017; 3(4): 524–548.
2. Torre LA, Bray F, Siegel RL et al. Global cancer statistics, 2012. *CA: a Cancer Journal for Clinicians* 2015; 65(2): 87–108.
3. Richards CH, Leitch FE, Horgan PG et al. A systematic review of possum and its related models as predictors of post-operative mortality and morbidity in patients undergoing surgery for colorectal cancer. *Journal of Gastrointestinal Surgery* 2010; 14(10): 1511–1520.
4. Jorgensen M, Young J, Dobbins T et al. A mortality risk prediction model for older adults with lymph node-positive colon cancer. *European Journal of Cancer Care* 2015; 24(2): 179–188.
5. Manilich EA, Kiran RP, Radivoyevitch T et al. A novel data-driven prognostic model for staging of colorectal cancer. *Journal of the American College of Surgeons* 2011; 213(5): 579–588.
6. Goos JA, Coupé VM, van de Wiel MA et al. A prognostic classifier for patients with colorectal cancer liver metastasis, based on aurka, ptgs2 and mmp9. *Oncotarget* 2016; 7(2): 2123–2134.
7. Quintana JM, Antón-Ladislao A, González N et al. Predictors of one and two years' mortality in patients with colon cancer: A prospective cohort study. *PloS one* 2018; 16(6): e0199894. DOI:https://doi.org/10.1371/journal.pone.0199894.
8. Berger AC, Sigurdson ER, LeVoyer T et al. Colon cancer survival is associated with decreasing ratio of metastatic to examined lymph nodes. *Journal of Clinical Oncology* 2005; 23(34):

8706–8712.

9. De Ridder M, Vinh-Hung V, Van Nieuwenhove Y et al. Prognostic value of the lymph node ratio in node positive colon cancer. *Gut* 2006; 55(11): 1681–1681.
10. Rosenberg R, Friederichs J, Schuster T et al. Prognosis of patients with colorectal cancer is associated with lymph node ratio: a single-center analysis of 3026 patients over a 25-year time period. *Annals of Surgery* 2008; 248(6): 968–978.
11. Rosenberg R, Engel J, Bruns C et al. The prognostic value of lymph node ratio in a population-based collective of colorectal cancer patients. *Annals of Surgery* 2010; 251(6): 1070–1078.
12. Edge SB, Byrd DR, Carducci MA et al. *AJCC cancer staging manual*, volume 7. Springer New York, 2010.
13. Quintana JM, Gonzalez N, Anton-Ladislao A et al. Colorectal cancer health services research study protocol: the CCR-CARESS observational prospective cohort project. *BMC Cancer* 2016; 16(1): 435.
14. Barrio I, Arostegui I, Rodríguez-Álvarez MX et al. A new approach to categorising continuous variables in prediction models: Proposal and validation. *Statistical Methods in Medical Research* 2017; 26(6): 2586–2602.
15. Barrio I, Rodríguez-Álvarez MX, Meira-Machado L et al. Comparison of two discrimination indexes in the categorisation of continuous predictors in time-to-event studies. *SORT* 2017; 1: 73–92.
16. Altman D. *Encyclopedia of Biostatistics*, chapter Categorizing Continuous Variables. John Wiley and Sons, Ltd. ISBN 9780470011812, 2005.

17. Arostegui I, Gonzalez N, Fernández-de Larrea N et al. Combining statistical techniques to predict postsurgical risk of 1-year mortality for patients with colon cancer. *Clinical Epidemiology* 2018; 10: 235–251.

18. McCullagh P and Nelder J. *Generalized Linear Models, 2nd ed*. London: Chapman & Hall, 1989.

19. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; 12: 387–415.

20. Hanley JA and McNeil BJ. The meaning and use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982; 143: 29–36.

21. Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2003.

22. Mebane W and Sekhon J. Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software* 2011; 42: 1–26.

23. Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 1979; 7(1): 1–26.

24. Roca-Pardiñas J, Cadarso-Suárez C, Nácher V et al. Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. *Statistics in Medicine* 2006; 25(14): 2483–2501.

25. Roca-Pardiñas J, Cadarso-Suárez C, Tahoces PG et al. Selecting variables in non-parametric regression models for binary response an application to the computerized detection of breast cancer. *Statistics in Medicine* 2009; 28(2): 240–259.

26. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2018. URL http://www.R-project.org/.

27. Tsuruta H and Bax L. Polychotomization of continuous variables in regression models based on the overall c index. *BMC Medical Informatics and Decision Making* 2006; 6: 41.

28. Mazumdar M and Glassman J. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine* 2000; 19: 113–132.

29. Barrio I, Arostegui I, Quintana J et al. Use of generalised additive models to categorise continuous variables in clinical prediction. *BMC Medical Research Methodology* 2013; 13: 83.

30. Zhou S and Shen X. Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association* 2001; 96(453): 247–259.

31. Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 2002; 11(4): 735–757.

32. Valenzuela O, Delgado-Marquez B and Pasadas M. Evolutionary computation for optimal knots allocation in smoothing splines. *Applied Mathematical Modelling* 2013; 37(8): 5851–5863.

33. Harrell F, Califf R, Pryor D et al. Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association* 1982; 247: 2543–2546.

34. Harrell F, Lee K, Califf R et al. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; 3(2): 143–152.

35. Royston P and Altman D. Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine* 2010; 29(24): 2508–2520.