journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Supervised discretization can discover risk groups in cancer survival analysis

Iván Gómez [a,b,*], Nuria Ribelles [b,c], Leonardo Franco [a,b], Emilio Alba [b,c], José M. Jerez [a,b]

[a] Computer Science Department, University of Málaga, Campus de Teatinos S/N, 29071 Málaga, Spain
[b] Málaga Biomedical Research Institute (IBIMA), Málaga, Spain
[c] Virgen de la Victoria Oncology Service, Málaga, Campus de Teatinos S/N, 29071 Málaga, Spain

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form
7 July 2016
Accepted 12 August 2016

Keywords:
CAIM
Decision Trees
ChiMerge
TNM protocol
Breast cancer free survival
Predictive models

ABSTRACT

Discretization of continuous variables is a common practice in medical research to identify risk patient groups. This work compares the performance of gold-standard categorization procedures (TNM+A protocol) with that of three supervised discretization methods from Machine Learning (CAIM, ChiM and DTree) in the stratification of patients with breast cancer. The performance for the discretization algorithms was evaluated based on the results obtained after applying standard survival analysis procedures such as Kaplan–Meier curves, Cox regression and predictive modelling. The results show that the application of alternative discretization algorithms could lead the clinicians to get valuable information for the diagnosis and outcome of the disease. Patient data were collected from the Medical Oncology Service of the Hospital Clínico Universitario (Málaga, Spain) considering a follow up period from 1982 to 2008.

1. Introduction

Decisions about how to treat breast cancer patients after surgery have been contingent on the accuracy of estimating the behaviour and outcome of the disease. Prognostic factors are capable of providing information about the clinical outcome independently of the therapy, and are usually indicators of growth, invasion, and metastatic potential [1]. For example, TNM staging of malignant tumour is derived from the biological progression of tumour growth: a solid tumour first grows locally (T), metastases to lymph nodes may occur (N), and finally other organs may be affected by metastases (M). In the case of breast cancer, tumour size is one of the most powerful predictors of its behaviour [2,3], whereas axillary lymph node status has repeatedly been shown to be the single most important predictor of disease-free survival and overall survival in breast cancer [4,5]. Also, and together with TNM classification, other features, such as age at diagnosis, have been identified as relevant prognostic factor for disease outcome.

Nevertheless, it is not clearly defined what is the optimal cutoff for selecting prognostic groups [6,7]. In fact, the TNM staging system was initially developed in 1950 and after successive amendments still remains the main prognostic

* *Corresponding author.* Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Campus de Teatinos S/N, Málaga 29071, Spain. Fax: +34 952131397.
E-mail address: ivan@lcc.uma.es (I. Gómez).

http://dx.doi.org/10.1016/j.cmpb.2016.08.006

classification in many adult cancers in order to predict treatment outcomes [3,8,9]. However, nowadays in the era of personalized medicine, a wide consensus exists about the need to exploit information more efficiently. For example, although the different classes of tumour size and the number of affected axillary lymph nodes have an established prognostic value in breast cancer, it may be possible to define other cutoffs that provide more prognostic information than those classically established. In this sense, the use of discretization algorithms from machine learning could help define these new classes of tumour size, axillary nodal status, or age at diagnoses. Better stratification of the patients into prognostic groups is essential in daily practice not only to decide the best treatment option but also to design clinical trials or to compare the outcomes across population groups.

Typically, discretization algorithms have been applied as a pre-processing phase for machine learning methods that handle only discrete values, affecting drastically the prediction of the classifiers, because usually more states of covariates require more learning effort. In the biomedical context where data frequently have high dimensionality, the discretization process can significantly improve the performance of classifiers [10,11]. This work analyses the performance in the application of common discretization methods to prognostic factors in the breast cancer survival context. Specifically, the authors focus on four supervised discretization methods, TNM+A (TNM+Age) [12], CAIM [13], ChiMerge [14], and Decision Trees [15], and their ability to find novel risk groups when classical survival analysis procedures are applied to clinical data (Kaplan–Meier estimator, proportional hazard regression models (Cox), and outcome predictive models).

This paper is organized as follows. Section 2 contains a description of the patient database included in this experiment with a short description of the specific discretization algorithms chosen to evaluate its performance. Section 3 represents the experimental results in a survival analysis considering the data segmented by the discretization methods. Lastly, Section 4 provides a short discussion and future directions for research.

2. Methods

2.1. Patient data

Data from 1248 patients with breast cancer disease from the Medical Oncology Service of the Hóspital Clínico Universitario (Málaga, Spain) were collected and recorded considering a follow-up period from 1982 to 2008. Clinico-pathological and follow-up information were obtained by chart review. Patients who died from any other cause were censored at the time of death. Eight prognostic variables considered by the physicians as very significant prognostic markers were selected and included in this study: age, tumour size, axillary lymph nodes, tumour histological grade, menopausal status, administration of adjuvant chemotherapy, use of adjuvant hormone therapy, and intrinsic subtype defined by four biomarkers (phenotype). Survival time and status completed the working dataset. Table 1 shows the ranges, statistics (mean or mode), and type of the covariates.

2.2. Discretization algorithms

The practice of discretizing continuous covariates (such as age or blood pressure) is common in medical and epidemiological research. It makes the analysis and interpretation of results simple and allows the clinicians to discover risk groups in clinical practice. Thus, categorized covariates may be preferred for offering a simpler interpretation of common effect measures from statistical models, such as odds ratios and relative risks or making data summarization more efficient [11,16–19].

Nevertheless, discretizing clearly implies some loss of information and searching for the optimal cut points for continuous variables. In fact, the optimal cutoff for selecting prognostic groups [6,7] is not clearly defined and, although some prognostic factors have classical established values (such as the TNM staging system in cancer), finding different cut points could provide valuable additional information. Thus, the need to discover different groups of patients for clinical decision making would justify the application of methodologies to find the optimal cut-points in continuous covariates [20].

Three common supervised discretization algorithms, CAIM, ChiMerge (ChiM) and Decision Trees (DTree), were analysed versus the standardized staging notation system in breast cancer. The comparison was done in terms of the results derived from the application of commonly used procedures in survival analysis: the Kaplan–Meier estimator, a Cox regression analysis, and outcome predictive models. The discretization [21] and hier.part [22] packages were used in R to apply the CAIM, ChiM and DTree algorithms.

2.2.1. The CAIM algorithm

One of the most representative supervised discretization algorithms is CAIM (Class–Attribute Interdependency

Table 1 – Range, mean or mode, and type of covariate are shown for the eight prognostic factors considered relevant by clinicians in breast cancer survival analysis.

Prognostic factors	Range	Mean or mode	Scale/Type
Age	25–88	56	Quantitative/Ratio
Tumour size	0.2–12	2.46	Quantitative/Ratio
Axillary lymph nodes	0–34	2.26	Quantitative/Ratio
Menopausal status	0, 1	1	Qualitative/Nominal
Histological grade	1, 2, 3	2	Qualitative/Nominal
Chemotherapy	0, 1	1	Qualitative/Nominal
Hormonal therapy	0, 1	1	Qualitative/Nominal
Phenotype	1, 2, 3, 4	1	Qualitative/Nominal
Class	0, 1	0	Qualitative/Nominal

Maximization Algorithm) [13]. The goal of this algorithm is to maximize the class–attribute interdependence and to generate a (possibly) minimal number of discrete intervals. It is included in the top–down category of methods, and compared with other discretization techniques, experiments have shown that it can generate an optimal discretization scheme.

The CAIM algorithm maximizes class-attribute interdependence by exploring a series of discretization points, to choose those that optimize a heuristic measure related to the dominance of a given class in the created intervals, including also a factor to minimize the number of intervals. It is a local algorithm that considers the input attributes independently, usually generating for each input variable a discrete representation with a length equal to the number of output classes (e.g., twice the original number of attributes for a binary output) [13,23]. The algorithm has been extensively tested leading to very good results and in several cases leading to classification rates larger than those obtained with the original data, a fact that can be due to a noise-filtering process produced as a side-effect in the discretization process.

2.2.2. *The ChiM algorithm*

ChiM is a simple algorithm that uses the chi-square statistic to categorize continuous attributes. It is a supervised, bottom-up data discretization method. ChiM is an improvement over the most obvious simple methods, such as equal-width intervals and equal-frequency intervals. It consists of an initialization step and a bottom-up merging process of intervals if a condition is met. The algorithm sorts the training examples according to their value for the attribute being discretized constructing the initial set of intervals, where each example is put into its own interval. Then, an interval merging process occurs with two steps: (1) compute the χ^2 value for each pair of adjacent intervals, (2) combine the pair of adjacent intervals with the lowest χ^2 value. This process continues until all pairs of adjacent intervals have a χ^2 value exceeding a threshold [14,24].

2.2.3. *DTree techniques*

The class of DTree techniques is a class of non-parametric regression trees applicable to all regression problems, including nominal, ordinal, numeric, censored, multivariate response variables, and arbitrary measurement scales of the covariates. DTree techniques estimate a regression relationship by binary recursive partitioning in a decision framework. This algorithm tests the null hypothesis of independence between the covariates and the response. If this hypothesis cannot be rejected, the algorithm stops. Otherwise, it selects the covariate with the strongest association to the response and splits it, repeating this process recursively [25].

2.3. *Statistical and computational survival analysis procedures*

The survival package in R was used to compute survival estimates and perform a multivariate regression analysis [26]. Clinico-pathological covariates were analysed and breast cancer free survival (BCFS), defined as the time from surgery until a recurrence or death from breast cancer, was considered as the clinical endpoint in the survival outcomes. Firstly, an actuarial survival test was performed by the Kaplan–Meier method and the significance in statistical differences was assessed using the log-rank test. Then, a Cox proportional hazard regression model [27] was used to examine the relationships of BCFS and the prognostic factors, and all possible combinations of covariates were tested to identify the best model according to the Akaike Information Criterion (AIC) [28]. For Cox models, the assumption of hazard proportionality was checked by testing for non-slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time [29].

Finally, a predictive model of recurrence in breast cancer was estimated by using a naive Bayes classifier. We have selected a naive Bayesian approach due to its fast implementation and simplicity. In addition, although the assumption of all attribute independence of each other is rarely true in the most real-world situations, the naive Bayesian classifier performs very well even in these cases [30]. This simple model will predict directly the probability of relapse within 36 months after surgical intervention, considering survival for a fixed time period and giving thus a classification problem. The naive Bayes classifier is a standard model which classifies new samples into the most probable class based on posterior probability computed according to the Bayes theorem. We also employed a support vector machine as learning method, but we rejected it because we got the same results but 20% slower performance. Even if initially Artificial Neural Networks (ANN) were also potentially considered as prediction method, we have discarded its use due to the difficulties involved in finding a suitable set of parameter adjustment, like number of layers and nodes. The area under the Receiver Operating Characteristic curve (AUC) was used as the accuracy metric to measure the predictive ability of the classification model [31]. A ten-fold cross validation scheme was used as the re-sampling procedure to do an internal validation. With this procedure, the data are randomly divided into ten subsets, then each one in turn is retained to test the model while the remaining nine subsets are used as training data. At the end, the partial test results are averaged to produce a single estimate of the global AUC.

3. Results

Different subgroups of patients were constituted after applying the CAIM, ChiM, and DTree discretization algorithms. Table 2 shows the cut-points and number of patients (N) in each subgroup. Also, the standardized staging system for tumour size plus affected lymph nodes and age (labelled by TNM+A) was added to the first column in the table to make further comparisons easy.

At first glance, the intervals for the selected patient subgroups differ depending on the algorithm applied to the covariates. Therefore, an exhaustive analysis was subsequently conducted to identify the benefits for the survival analysis when three classical procedures (Kaplan–Meier estimators, Cox proportional hazard models, and predictive models) were applied to the patient dataset.

Table 2 – Cut-points obtained by the application of TNM+A protocol, CAIM, ChiM and Decision Tree techniques.

Variables	TNM+A		CAIM		ChiM		DTree	
Age	Cut-points	N	Cut-points	N	Cut-points	N	Cut-points	N
	<40	111	<33.5	39	<33.5	39	≤35	46
	40–55	468	≥33.5	1209	33.5–57.5	608	≥35	1202
	≥55	669			≥57.5	601		
Size	<2	496	<2.35	686	<0.55	14	≤1.8	426
	2–5	666	≥2.35	540	0.55–1.45	254	1.8–4	679
	≥5	64			1.45–2.35	418	≥4	121
					≥2.35	540		
Lymph nodes	0	638	≤9.5	1156	<0.5	638	≤1	819
	1–3	349	>9.5	84	0.5–3.5	349	1–3	168
	≥4	253			3.5–11.5	199	≥3	253
					≥11.5	54		

3.1. *The Kaplan–Meier estimator*

The Kaplan–Meier (KM) plots and log-rank values generated by every discretization method along with the TNM+A staging system to the factors age, tumour size, and number of lymph nodes are depicted in Figs. 1–3.

The survival curves obtained by applying the CAIM, ChiM and DTree algorithms are slightly different from those coming from the TNM+A classification. More precisely, regarding patient age, Fig. 1 reveals a different progression between the youngest groups discovered for the three algorithms (less than 33.5 years in CAIM and ChiM, 35 years in DTree) and the others. Besides, the *p*-value associated with the log-rank test for comparing survival curve also decreases, which means there is a more significant statistical difference between the patient groups. TNM+A traditionally established the low risk category as having age less than 40 years, so this interesting result could lead clinicians to question the potential benefits of established treatments for patients in this group.

In the case of tumour size, the KM-plots (Fig. 2) revealed very similar patient progressions for the TNM+A and DTree algorithms. The same situation is observed in the CAIM and ChiM

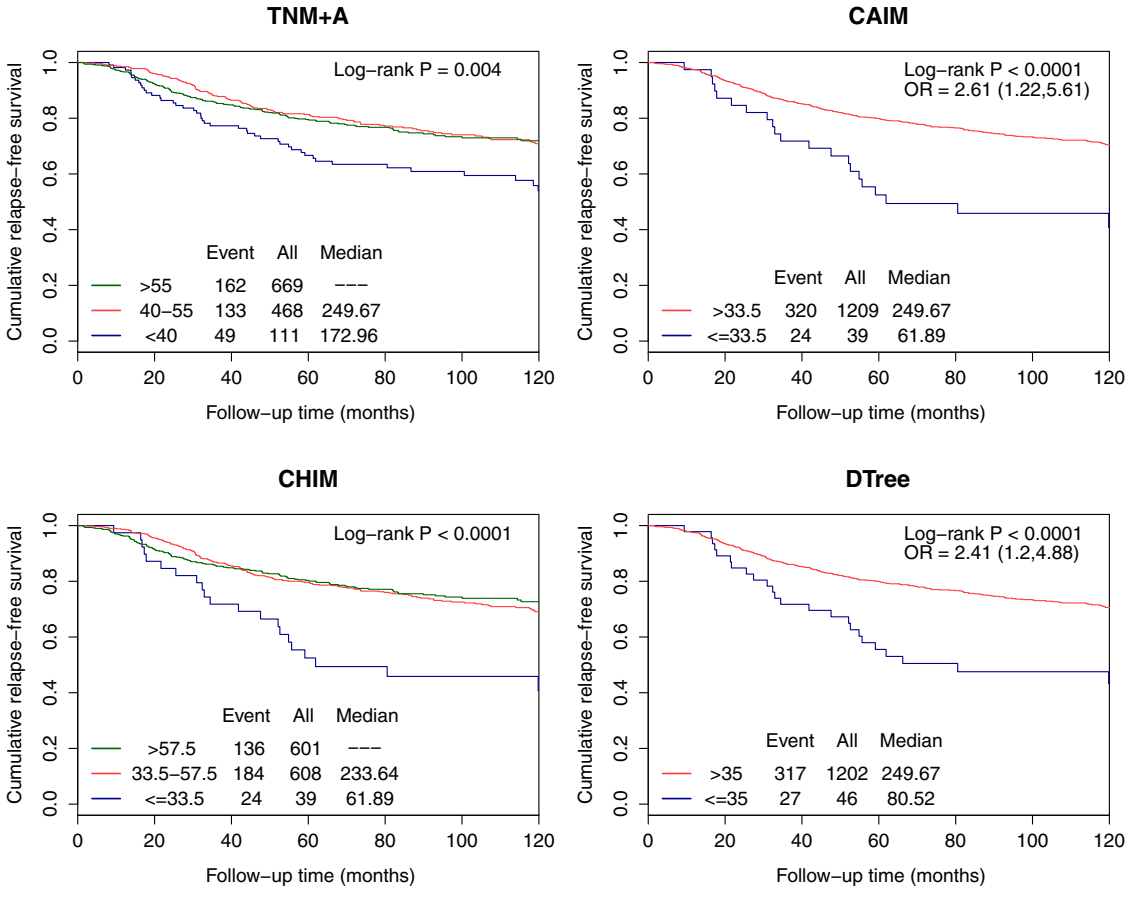

Fig. 1 – KM-plots for *age* covariate.

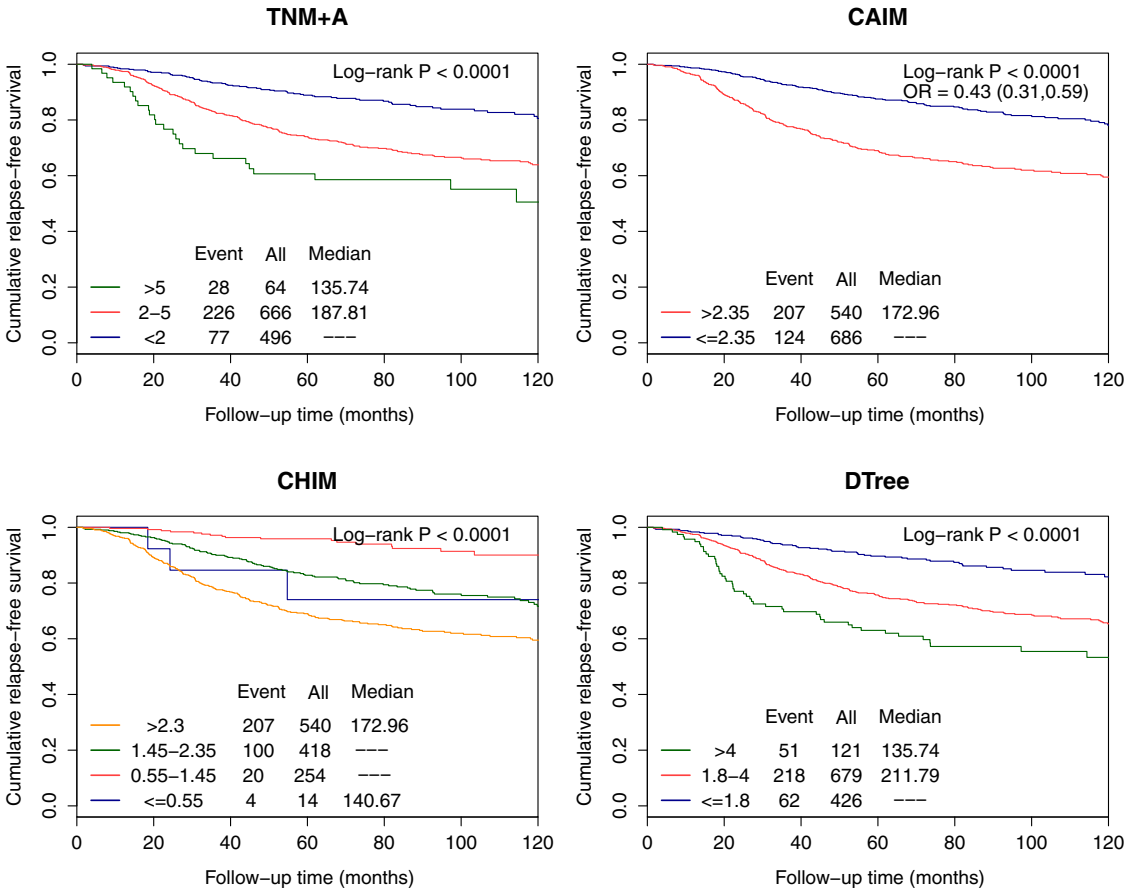

Fig. 2 – KM-plots for *tumour size* covariate.

plots. In fact, the log-rank test shows the same associated *p*-value for every method. Nevertheless, a lower risk group (≤0.55) is found by the ChiM algorithm; therefore, further analysis could be done by clinicians to study differences in risk between patients with tumour sizes in the intervals less than 0.55 cm and 0.55–2 cm. Moreover, the result obtained by the CAIM algorithm indicates that it is probably not necessary to split the tumour size into three different categories (as done by the TNM protocol), but separating into two groups clearly allows keeping the differences in survival without loss of statistical power. In addition, keeping two groups usually increases their population sizes, which is important for further statistical analysis.

With regard to the covariate number of lymph nodes, Fig. 3 shows similar KM-plots from the application of the TNM+A and Dtree methods. However, the ChiM and CAIM algorithms clearly find other risk subgroups for patients with numbers of nodes beyond the threshold of 4 cm established by the TNM protocol. Indeed, the KM curves reflect a poor progression of patients with more than 9.5 and 11.5 affected lymph nodes for the CAIM and ChiM algorithms, respectively.

3.2. *Proportional Cox model*

The Cox model is a well-recognized and popular statistical technique used in survival analysis [32]. It allows analysing the effect of several risk factors on survival times of individuals through the hazard function. It is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and a set of explanatory variables. This hazard function may be written as a multiplicative model,

$$h_i(t) = h_0(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}).$$

Here, *i* is a subscript for observation and *x* are the covariates. The constant $h_0(t)$ in this model represents a kind of log-baseline hazard.

The **Attribute** column in Table 3 shows the covariates of the best multivariate Cox model of BCFS itemized by algoritm. The hazard ratio (**HR** column in Table 3) can be interpreted as the chance of an event occurring in a covariate group divided by the chance that it happens in the control group. It corresponds to hazard ratio compared to the control value of each explanatory variable of the Cox model. The **CI** column indicates the confidence interval and **p** is the associated *p*-value.

Cox proportional hazard models were estimated to analyse the association between BFCS and the selected prognostic factors. An exhaustive search for the best model (for each discretization algorithm) was made by testing all the possible combinations of the covariates. The assumption of hazard proportionality was checked for every model analysed through the scaled Schoenfeld residuals.

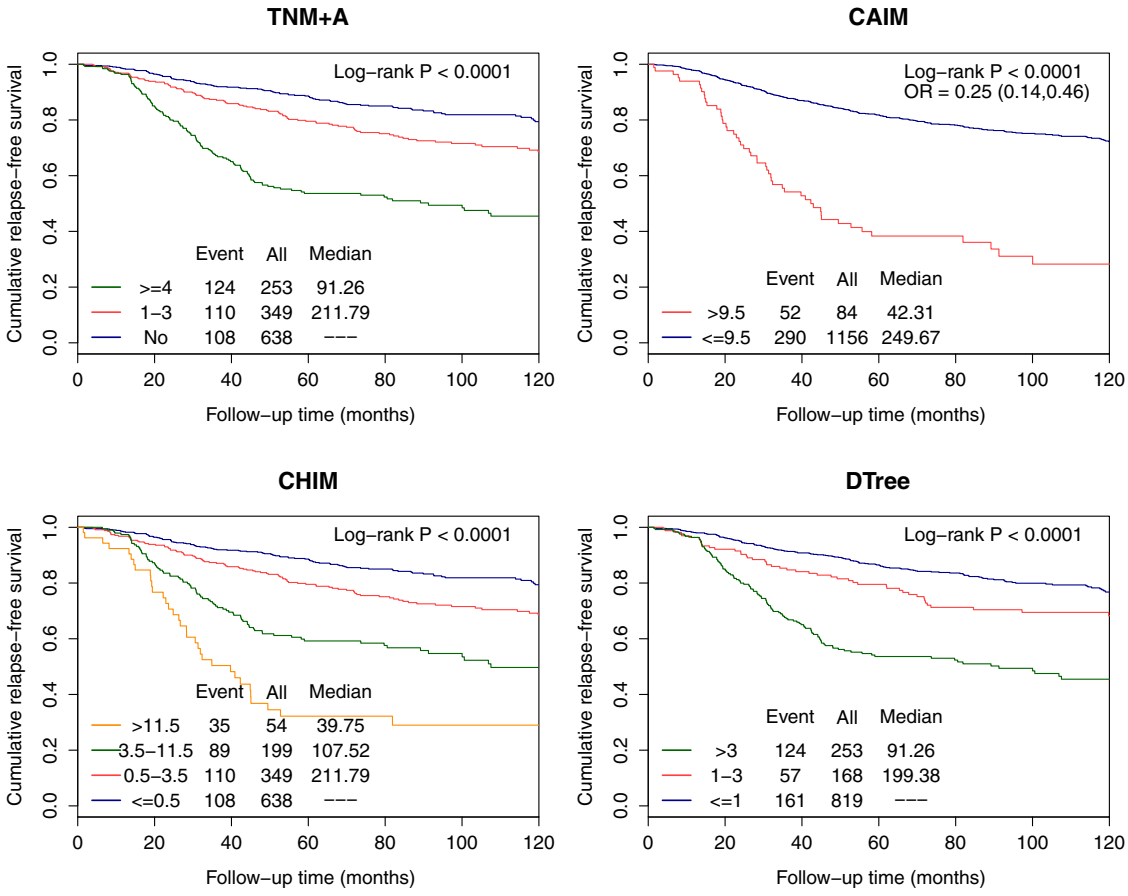

Fig. 3 – KM-plots for *number of lymph nodes* covariate.

The results in Table 3 show that the covariate *number of lymph nodes* remains in all the final models, which is not surprising, since its importance in predicting a relapse in breast cancer has been well documented in the literature. Nevertheless, a difference between the hazard ratio associated with this prognostic factor seems to be significant for the CAIM and ChiM algorithms, with hazard ratios of 3.88 and 9.42 respectively in the highest risk subgroups, compared to the

Table 3 – Best multivariate Cox model BCFS itemized by algoritm. The HR column corresponds to hazard ratio compared to the control value of the explanatory variable. The CI column indicates the confidence interval and p is the associated *p*-value.

	Attribute	Value	HR	CI (95%)	p	
TNM+A	**Age**	40–55	0.65	0.45	0.93	0.019
		>55	0.69	0.49	0.99	0.043
	Size	2–5	1.85	1.39	2.45	1.73*e*–05
		>5	2.74	1.72	4.36	1.98*e*–05
	Nodes	1–3	1.59	1.19	2.13	0.001
		≥ 4	3.20	2.39	4.28	4.00*e*–15
CAIM	**Age**	>33.5	0.49	0.30	0.80	0
	Men.State	*Postmen*	0.88	0.69	1.12	0.31
	Size	>2.35	2.19	1.73	2.78	9.14*e*–11
	Nodes	>9.5	3.88	2.75	5.46	8.10*e*–15
ChiM	**Nodes**	0.5–3.5	1.82	1.32	2.4	3.45*e*–05
		3.5–11.5	3.48	2.59	4.68	<2*e*–16
		>11.5	9.42	5.83	15.22	<2*e*–16
DTree	**Age**	>35	0.47	0.22	–3.30	0.0027
	Nodes	>3	1.8	0.07	8.58	<2*e*–16
	Chemo	ANTRA–TAX	0.66	0.18	–2.16	0.03
		CMF	0.99	0.16	0	0.99
		NO	0.90	0.15	–0.63	0.52
		TAX	0.68	1.00	–0.37	0.71

Table 4 – Reclassification of patients according to time of recurrence in a period of 36 months from surgery. t_f = time of follow-up from surgery.

$$status = \begin{cases} 1, & t_f <= 36 \,\&\, relapse = true \\ 0, & t_f <= 36 \,\&\, relapse = false \,||\, t_f > 36 \\ deleted, & t_f <= 36 \,\&\, censored = true \end{cases}$$

hazard ratio of 3.2 in the TNM+A classification scheme. This result is coherent with the previously described survival curves in the KM procedure. More specifically, it is possible to observe that the number of lymph nodes is the only prognostic factor that remains in the best model for the ChiM algorithm. This can lead one to think that this splitting of the risk subgroups confers a more significant importance to this covariate as a prognostic factor in breast cancer relapse, than that established by the traditional TNM+A protocol.

3.3. *Relapse prediction model*

In this section, survival analysis is formulated as a classification problem (relapse, non-relapse) by building a naive Bayes model that provides the probability of the event of interest's happening within 36 months after surgery [33,34] (fixed-time survival models). Machine Learning algorithms have been widely used in survival analysis due to their ability to find complex interactions in the data, in comparison to traditional statistical procedures [34–36].

In fixed-time prediction models, the cohort of patients is re-classified according to the time to recurrence. Thus, on the one hand, patients who relapsed before 36 months from surgery are assigned to a new status, "relapse", whereas those patients censored in this interval were removed from the analysis because it is not possible to know whether a recidivism would have occurred for them before 36 months. On the other hand, patients whose last follow-up time was greater than 36 months were assigned to a new status, "non-relapse", independently of whether they suffered recidivism or not after 36 months of follow-up (Table 4).

In order to compare the influence of the different discretization strategies in the precision accuracy of the classifier, exhaustive simulations to find the best prediction model were done for the CAIM, ChiM and DTree algorithms. Moreover, the best model, including the covariates of interest in their continuous form (without discretization, labelled as *NODIS*), was explored to complete the analysis. The remaining prognostic factors (menopausal status, histological grade, chemo and hormonal therapy, and phenotype) were included in their original form.

Table 5 illustrates the means and standard deviations for AUC averaged over a ten-fold cross validation procedure. The same results are depicted in Fig. 4, which shows similar distributions of the AUC for the TNM+A, DTree and CAIM algorithms, whereas the ChiM algorithm clearly outperforms the others, including *NODIS*, in terms of prediction ability. The statistical significance of the differences in AUC observed were analysed according to Ref. [37], where the author proposes using the Friedman test [38] on averaged results when a cross-validation procedure is used as a re-sampling method. The Friedman test is a non-parametric test (similar to an ANOVA test) that compares the average ranks of K methods ($K > 2$).

Table 5 – Summary statistics for the classifier results.

	CAIM	DTree	TNM+A	ChiM	NODIS
Mean	0.7081	0.7109	0.7202	0.7519	0.7422
Standard deviation	0.0832	0.0811	0.0701	0.0587	0.0495

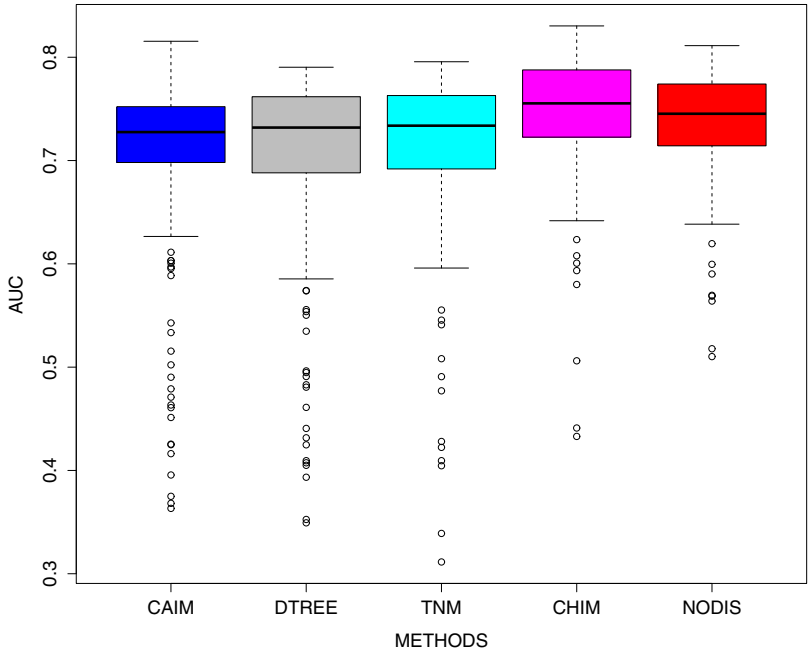

Fig. 4 – Classification accuracy results performed by a naive Bayesian classifier.

Table 6 – Nemenyi test between original and discretized data.

	CAIM	DTree	TNM+A	ChiM
DTree	0.0366	—	—	—
TNM+A	0.9797	0.0059	—	—
ChiM	≤2e–16	≤2e–16	≤2e–16	—
NODIS	≤2e–16	≤2e–16	1.1e–14	2.3e–05

Under the null hypothesis, the Friedman test states that all algorithms perform an equivalently way, the observed differences being merely random. If the null hypothesis is rejected (the means are significantly different from each other), the Nemenyi test is used to discover which means are different from the others.

The application of the Friedman test identified a statistically significant difference in predicting breast cancer relapse within 3 years depending on which type of discretization algorithm is used ($\chi^2(4) = 567.76$ and $p \leq 2.2e-16$). The results of the Nemenyi post hoc test are depicted in Table 6, where the *p*-values corresponding to pairwise comparison of accuracy predictions between the methods are illustrated, also if no discretization is applied. Regarding the TNM+A method, the Nemenyi post hoc test found no statistical diference from the CAIM, and a slight difference from the Dtree algorithm.

In addition, the ChiM method differs, highly significantly, from the other analysed options. This may be caused by the large number of intervals generated by this algorithm and therefore less loss of information induced by the categorization. However, the ChiM method shows better results even if no discretization is applied to the covariates of interest. This interesting fact might be due to a good selection of cut-off points made by ChiM algorithm, which removes noisy elements from the original dataset.

4. Conclusions

The discretization of continuous variables is widely perceived as a common practice to simplify the analysis in clinical research. Although this procedure may lead to some information loss, the predictive relevance of categorical variables is more easily established than their continuous form, improving the model interpretability.

However, there is no clear definition of the optimal method of establishing the cut-off values to define these groups. The choice of cut-points is often based on physicians' experience and previous research results. Although some prognostic factors have classical established values, finding other, different, cut-points could provide useful further information.

In this paper, we have selected some of the most representative discretization techniques from Machine Learning to classify patients into groups regarding some prognostic factors. We have focused on clarifying which is the best method to obtain the boundaries that define these prognostic groups. Subsequently, we compared the results with the TNM+A protocol used as a standard in clinical practice.

The Kaplan–Meier curves generated by the CAIM, ChiM and DTree methods were slightly different for the number of axillary nodes and tumour size prognostic factors. However, we found an interesting fact regarding the age covariate, because each discretization method found a prognostic group younger than that of the standard TNM+A method, with a substantial improvement in the statistical significance.

The importance of the covariate number of lymph nodes in predicting relapse in breast cancer is confirmed, as it remains in the final model whichever method is used. Specifically, the hazard ratios associated to the highest groups give an important statistical value to this covariate.

One of the most important findings of this paper is that using machine learning techniques to find new prognostic groups of breast cancer patients could help improve the prediction of relapse. The extensive simulations performed showed that the application of these techniques to segment the data leads to efficient results in prediction accuracy. In fact, the ChiM algorithm exhibited the best statistical result for breast cancer prediction relapse within 36 months after surgery.

Finally as an overall conclusion, we can say that our analysis has found strong evidence suggesting that the standard TNM classification system can be easily improved by modifying the cut-off points and by the addition of extra variables. These changes will create new patient groups improving the decision making process in relation to choosing the best treatment, decreasing the cases of both overtreatment and undertreatment, thus contributing to the advancement of personalized medicine.

REFERENCES

[1] G. Gasparini, F. Pozza, A.L. Harris, Evaluating the potential usefulness of new prognostic and predictive indicators on node-negative breast cancer patients, J. Natl. Cancer Inst. 85 (15) (1993) 1206–1219.

[2] E. Fisher, S. Land, B. Fisher, L. Middleton, Pathologic findings from the national surgical adjuvant breast and bowel project: twelve-year observations concerning lobular carcinoma in situ, J. Breast Dis. 15 (3) (2004) 274–275.

[3] R. Rami-Porta, D. Ball, J. Crowley, D. Giroux, J. Jett, W. Travis, et al., The IASLC lung cancer staging project: proposals for the revision of the T descriptors in the forthcoming (seventh) edition of the TNM classification for lung cancer, J. Thorac. Oncol. 2 (2007) 593–602.

[4] E.R. Fisher, S. Anderson, C. Redmond, B. Fisher, Pathologic findings from the national surgical adjuvant breast project protocol B-06 10-year pathologic and clinical prognostic discriminants, Cancer 71 (8) (1993) 2507–2514.

[5] J.D. Seidman, L.A. Schnaper, S.C. Aisner, Relationship of the size of the invasive component of the primary breast carcinoma to axillary lymph node metastasis, Cancer 75 (1) (1995) 65–71.

[6] J.L. Gnerlich, A.D. Deshpande, D.B. Jeffe, A. Sweet, N. White, J.A. Margenthaler, Elevated breast cancer mortality in women younger than age 40 years compared with older women is attributed to poorer survival in early-stage disease, J. Am. Coll. Surg. 208 (3) (2009) 341–347.

[7] A.M. Gonzalez-Angulo, K. Broglio, S.-W. Kau, Y. Eralp, J. Erlichman, V. Valero, et al., Women age ≤ 35 years with primary breast carcinoma, Cancer 103 (12) (2005) 2466–2472.

[8] F.L. Greene, L.H. Sobin, A worldwide approach to the TNM staging system: collaborative efforts of the AJCC and UICC, J. Surg. Oncol. 99 (5) (2009) 269–272.

[9] Y. Hashiguchi, K. Hase, K. Kotake, H. Ueno, E. Shinto, H. Mochizuki, et al., Evaluation of the seventh edition of the

tumour, node, metastasis (TNM) classification for colon cancer in two nationwide registries of the United States and Japan, Colorectal Dis. 14 (9) (2012) 1065–1074.
[10] J.L. Lustgarten, V. Gopalakrishnan, H. Grover, S. Visweswaran, Improving classification performance with discretization on biomedical datasets, in: AMIA Annual Symposium Proceedings /AMIA Symposium, 2008, pp. 445–449.
[11] D.M. Maslove, T. Podchiyska, H.J. Lowe, Discretization of continuous features in clinical datasets, J. Am. Med. Inform. Assoc. 20 (3) (2013) 544–553.
[12] S.E. Singletary, F.L. Greene, Revision of breast cancer staging: the 6th edition of the TNM classification, Semin. Surg. Oncol. 21 (1) (2003) 53–59.
[13] L.A. Kurgan, K.J. Cios, CAIM discretization algorithm, IEEE T. Knowl. Data En. 16 (2) (2004) 145–153.
[14] H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, in: Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995, pp. 388–391.
[15] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[16] A.B. Cantor, J.J. Shuster, Re: Dangers of using "optimal" cutpoints in the evaluation of prognostic factors, J. Natl. Cancer Inst. 86 (23) (1994) 1798–1799.
[17] G. Madghu, A new discretization method for continuous laboratory features in the diagnosis of dengue fever, Indian J. Med. Inform. 8 (1) (2014) 1–10.
[18] M. Mazumdar, A. Smith, J. Bacik, Methods for categorizing a prognostic variable in a multivariable setting, Stat. Med. 22 (2003) 559–571.
[19] B.A. Williams, J.N. Mandrekar, S.J. Mandrekar, S.S. Cha, A.F. Furth, Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes, Tech. rep., Department of Health Sciences Research, Mayo Clinic, Rochester, MN, June, 2006.
[20] B. Lausen, T. Hothorn, F. Bretz, M. Schumacher, Assessment of optimal selected prognostic factors, Biom. J. 46 (3) (2004) 364–374.
[21] H. Kim, Discretization: data preprocessing, discretization for classification, R package version 1.0-1, 2012.
[22] C. Walsh, R. Mac Nally, hier.part: hierarchical partitioning, R package version 1.0-4, 2013.
[23] C.-J. Tsai, C.-I. Lee, W.-P. Yang, A discretization algorithm based on class-attribute contingency coefficient, Inf. Sci. 178 (3) (2008) 714–731.
[24] R. Kerber, ChiMerge: discretization of numeric attributes, in: Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA, 1992, pp. 123–128.
[25] K.L. Salis, S. Kliem, K.D. O'Leary, Conditional inference trees: a method for predicting intimate partner violence, J. Marital Fam. Ther. 40 (4) (2014) 430–441.
[26] T.M. Therneau, A package for survival analysis in S, version 2.38, 2015.
[27] D.R. Cox, Regression models and life-tables, J. Roy. Stat. Soc. B 34 (2) (1972) 187–220.
[28] Y. Sakamoto, M. Ishiguro, G. Kitagawa, Akaike information criterion statistics, KTK Dordrecht, Tokyo, 1986.
[29] D. Schoenfeld, Partial residuals for the proportional hazards regression model, Biometrika 69 (1) (1982) 239–241.
[30] I. Rish, An empirical study of the naive Bayes classifier, Tech. rep., 2001.
[31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, Berlin, 2001.
[32] F.E. Harrell, Cox Proportional Hazards Regression Model, Springer New York, New York, 2001, pp. 465–507.
[33] R. Demicheli, P. Valagussa, G. Bonadonna, Does surgery modify growth kinetics of breast cancer micrometastases?, Br. J. Cancer 85 (4) (2001) 490.
[34] J. Jerez, L. Franco, E. Alba, A. Llombart-Cussac, A. Lluch, N. Ribelles, et al., Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks, Breast Cancer Res. Treat. 94 (3) (2005) 265–272.
[35] L. Vanneschi, A. Farinaccio, G. Mauri, M. Antoniotti, P. Provero, M. Giacobini, A comparison of machine learning techniques for survival prediction in breast cancer, BioData Min. 4 (1) (2011) 1–13.
[36] B. Zupan, J. Demšar, M.W. Kattan, J. Beck, I. Bratko, Machine learning for survival analysis: a case study on recurrence of prostate cancer, Artif. Intell. Med. 20 (1) (2000) 59–75, Selected Papers from {AIMDM} '99.
[37] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[38] M. Friedman, A comparison of alternative tests of significance for the problem of *m* rankings, Ann. Math. Statist. 11 (1) (1940) 86–92.