J. Vaillant[1], M. Troupé[1], J. Manuceau[1], V. Lánska[2]

Nonparametric Selection Method of Survival Predictors with an Application to Breast Cancer

[1]UFR Sciences, Université des Antilles-Guyane, Pointe-à-Pitre, Guadeloupe,
[2]Institute for Clinical and Experimental Medicine, Prague, Czech Republic

Abstract: Mutual information between a survival variable and covariables provides a new tool for selecting covariables with a high predictive value whenever there is no reasonable parametric model with respect to the observed phenomena. The information rate carried out by covariables can be tested by means of a decomposition similar to the analysis of variance. Moreover, a method based on information conservation can be used for aggregating survival curves corresponding to different modalities of the same selected predictor which increases the prediction efficiency. These results are applied to survival data from 1304 patients with breast cancer followed over a period of ten years.

Keywords: Entropy, Information, Predictor, Survival Analysis

1. Mathematical Background

Different types of measures are used in the field of information theory. Since the pioneer work published in [8] and [15], the literature on information measures has expanded and many diversity and divergence indices (related to entropy) are described by several authors (see, e. g.,[5, 11, 12, 14]). Other authors have recently discussed the application of information gain methods to quantitative epidemiology and survival analysis [4, 9]. In fact, the entropy of random variables is an interesting tool for characterising uncertainty associated with these variables.

For a random variable X on a countable set Ω, the entropy of X is

$$H\,(X) = -\sum_{x \in A} P\,(X = x) \ln P\,(X = x), \quad (1)$$

where A is the set of all possible values for X and $\left(P\,(X = x), x \in A\right)$ is its probability distribution.

If we now consider two random variables X and Y on Ω, the entropy of X conditional on Y is defined as follows:

$$H\,(X\,|\,Y) = -\sum_{(x,\,y) \in A \times B} P\,(X = x, Y = y) \ln P\,(X = x\,|\,Y = y) \quad (2)$$

where $P\,(X = x \mid Y = y)$ is the conditional probability of the event $X = x$ *given* $Y = y$; B is the set of all possible values for Y.

$H\,(X \mid Y)$ represents the average uncertainty on X when Y is known.

The notion of entropy seen above cannot be directly extended to noncountable sets (see [16] for the formal link between *Shannon* entropy [17] and *Kullback* entropy [8] and their respective properties).

Let μ be a measure [2] on X, the generalized entropy with respect to this reference measure μ ([2, 16]) is defined as:

$$H\,(X;\,\mu) = -\int_A f\,(x) \ln f\,(x)\,\mu\,(dx);$$

where f is the probability density of X with respect to μ. If X is singular with respect to μ, we set $H\,(X;\,\mu) = \infty$. It is worth noticing that if we consider the generalized entropy of the approximating discrete random variable X_n with respect to the corresponding discrete approximation μ_n to the reference measure μ, then $H\,(X;\,\mu)$ is the limit of the sequence of discrete approximations $H\,(X_n;\,\mu_n)$.

Let $X_1, \ldots, X_n$ be n random variables, we can write $X = (X_1, \ldots, X_n)$ and derive from the previous formulas the entropy of vector X. In particular, the entropy of a pair of random variables $(X,\,Y)$ is

$$\mathrm{H}\,(\mathrm{X},\,\mathrm{Y}) = -\sum_{(x,y) \in A \times B} P\,(X = x,\,Y = y) \ln P\,(X = x,\,Y = y). \quad (3)$$

The mutual information between X and Y is by definition

$$I\,(X,\,Y) = H\,(X)+H\,(Y)-H\,(X,\,Y) = H\,(X)-H\,(X|Y) \quad (4)$$

It can be easily checked that $0 \le I\,(X,\,Y) \le H\,(X)$, $I\,(X,\,X) = H\,(X)$ and $I\,(X,\,Y) = I\,(Y,\,X)$. On the other hand, $I\,(X,\,Y) = 0$ if and only if X and Y are independent. $I\,(X,\,Y)$ measures the average reduction in uncertainty about X by knowing Y and vice versa. Therefore, if Y is the variable of interest, for example the patient survival time after disease detection, and X a medical covariable, the relative information of X on Y defined as

$$I_r\,(X,\,Y) = \frac{I\,(X,\,Y)}{H\,(Y)} \quad (5)$$

is a nonlinear correlation coefficient taking values in [0, 1]. $I_r\,(X,\,Y)$ is the proportion of reduction in uncertainty about Y by knowing X.

2. Entropy Decomposition

We consider Y as the variable of interest, and $X_1,\dots, X_n$ its covariables (with $n > 1$). Following the definition of mutual information between Y and X_1, we can define the mutual information between Y and X_1 conditionally to a third variable, say X_2:

$$\begin{aligned} I(Y, X_1 \mid X_2) &= H(Y \mid X_2) + H(X_1 \mid X_2) - H\big((Y, X_1) \mid X_2\big) \qquad (6) \\ &= H(Y \mid X_2) - H(Y \mid X_1, X_2). \end{aligned}$$

$I(Y, X_1 \mid X_2)$ is the average reduction in uncertainty about Y when knowing both X_1 and X_2 instead of X_2 only.

We will now show how the entropy of Y can be decomposed by means of the mutual information between Y and the X_i.

Proposition 1 *(Entropy Decomposition). The entropy of Y can be decomposed as follows:*

$$\begin{aligned} H(Y) = &\sum_{i=1}^{n} I(Y, X_i) \\ &+ \sum_{i=2}^{n} \big(I((X_1, \dots, X_{i-1}), X_i \mid Y) \qquad (7) \\ &- I((X_1, \dots, X_{i-1}), X_i)\big) \\ &+ H\big(Y \mid (X_1, \dots, X_n)\big). \end{aligned}$$

This proposition is proved by induction on n in [9].

Expression (7) provides an entropy decomposition which parallels the square sum decomposition of the variance analysis. $H(Y \mid (X_1, \dots, X_n))$ can be considered as a residual term which is close to zero when the X_i carry out most of the information on Y.

When we have no reasonable parametric models with respect to the observed phenomena (i. e., variable of interest Y and $X_1, \dots, X_n$), a nonparametric approach is possible from the information provided by contingency tables associated with the observations in a similar way as the one used by [13] and [14].

3. Covariable Selection

We use a method based on the estimation of the conditional and nonconditional entropies of Y and covariables $X_1, \dots, X_n$. The estimates are calculated by means of the empirical frequencies provided by the multidimensional contingency tables. For covariables having a continuous distribution, the outcomes are grouped into discrete categories by applying an entropy concentration principle [16]. A data-dependent estimator for the mutual information can be obtained by means of a nested sequence of partitions made of rectangles as in [3].

Let Y be the variable of interest, X be any covariable, (N_{ij}) be the contingency table associated with Y and X, and $p = (p_{ij})$ be the matrix of expected relative frequencies. $p_{i.}$ and $p_{.j}$ refer to the marginal probabilities. The method of selection consists of quantifying the relationship between X and Y by using the relative information of X on Y since $I_r(X, Y)$ is the proportion of reduction in uncertainty about Y by knowing X. The observation of X and Y through a contingency table leads to the following property which is a consequence of equalities (1) and (5):

Property 1 *An expression of the relative information of X on Y is:*

$$I_r(X, Y) = \frac{\sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_{i.}p_{.j}}}{\sum_j p_{.j} \ln p_{.j}} \qquad (8)$$

We then deduce the following property:

Property 2 *An estimator of the relative information of X on Y is:*

$$T = \widehat{I_r}(X, Y) = \frac{\sum_{i,j} \widehat{p_{ij}} \ln \frac{\widehat{p_{ij}}}{\widehat{p_{i.}}\widehat{p_{.j}}}}{\sum_j \widehat{p_{.j}} \ln \widehat{p_{.j}}} \qquad (9)$$

where $\widehat{p_{ij}}$, $\widehat{p_{i.}}$ and $\widehat{p_{.j}}$ are classical notations used for the empirical relative frequencies.

This estimator is convergent and follows asymptotically the Gaussian distribution

$$N\big(I_r(X, Y), \text{var}(T)\big)$$

where var *(T) is the asymptotic variance of T. Its expression is*

$$\text{var}(T) = \frac{1}{N} \text{var}\Big(\sum_{i,j} \frac{\partial I_r(X, Y)}{\partial p_{ij}}(p)\, I_{ij}\Big) \qquad (10)$$

where $N = \sum_{ij} N_{ij}$ is the number of statistical units considered, and the I_{ij} are the indicator functions associated with the contingency table of (N_{ij}). The convergence considered above is as N becomes larger. Using equalities (8) and (10), we get the following property:

Property 3 *An expression of the asymptotic variance of estimator T is:*

$$\text{var}(T) = \frac{1}{N}\Big[\frac{1}{H^2(Y)} \sum_{i,j} p_{ij} \Big[\ln \frac{p_{ij}}{p_{i.}p_{.j}} + T \ln p_{.j}\Big]^2\Big]. (11)$$

var (T) can be estimated by $\widehat{\text{var}}(T)$ which is obtained by replacing p_{ij}, $p_{i.}$ and $p_{.j}$ by their empirical estimations $\hat{p}_{ij}$, $\hat{p}_{i.}$ and $\hat{p}_{.j}$.

In the case of information carried by several covariables, we can measure as above the combined relative information of covariables $X_1, \dots, X_n$ on Y. The evaluation of the information variance permits to build confidence intervals and eliminates the less pertinent covariables.

Vocabulary When Y is the survival variable, $I_r(X, Y)$ is called prognostic value of X and denoted by $I_r(X)$.

We select progressively the different covariables using an iterative process. Let E be the set of covariables $X_1, \dots, X_k$ considered in the study. The first step consists of building the set E_1 defined by

$$E_1 = \Big\{X_l \in E \mid \widehat{I_r}(X_l, Y) + \upsilon_\alpha \sqrt{\widehat{\text{var}}\big(\widehat{I_r}(X_l, Y)\big)} \geq p_1\Big\}$$

and at a step h ($h \geq 2$), we get

$$E_h = \Big\{(X_l, x) \in E \times E_{h-1} \mid \widehat{I_r}\big((X_l, x), Y\big) + \upsilon_\alpha \sqrt{\widehat{\text{var}}\big(\widehat{I_r}((X_l, x), Y)\big)} \geq p_h\Big\}$$

where p_1, p_h and α are fixed thresholds, and υ_α is such that $\phi(\upsilon_\alpha) = 1 - \frac{\alpha}{2}$; ϕ being the cumulative distribution function of the Gaussian distribution $N(0,1)$.

Remarks

1. The thresholds series$\{p_h, h \geq 1\}$ must be a strictly increasing sequence with respect to h because the adding of a covariable necessarily increases the relative information.
2. The selection process described above is interesting since at any given step h, a covariable which has not been selected at a previous step can be included in the selection set E_h.

In practice, the stopping criteria are the following:

i) a maximum number of iterations is determined by the user,
ii) the user determined a maximum threshold beyond which he considers that the relative information is sufficient,
iii) both *i)* and *ii)*,
iv) no covariables can be added.

4. Aggregation of Survival Curves

In the case where Y is the survival variable, the quotient of the number of patients alive at time t by the number of uncensored patients just before time t is called survival rate at time t. The survival curve is the curve of the survival rates as a function of t. The recovery rate of a sub-population is the survival rate at the end of the observation period.

Aggregating survival curves corresponding to different modalities of the same selected predictor increases the prediction efficiency if there is no loss of information due to this aggregation, and simplifies the prognostic procedure.

The different modalities of a covariable give a partitioning of the patient population *(cf. examples presented in Figs. 1 and 2)*. These modalities are ordered with respect to their recovery rates. If the survival curves of two modalities are very close, they can be grouped into a single modality. The method we apply is based on an entropy conservation principle presented in [10]: Aggregating two curves corresponding to two modalities of a covariable, say X, is equivalent to creating a new variable X'. This so-called modification of X has a number of modalities equal to the one of X minus unity. Then, a statistical test is applied to confirm if the relative information on the survival has not significantly decreased from X to X'. The decision rule consists of comparing the information loss with the appropriate percentile of a chi-squared distribution.

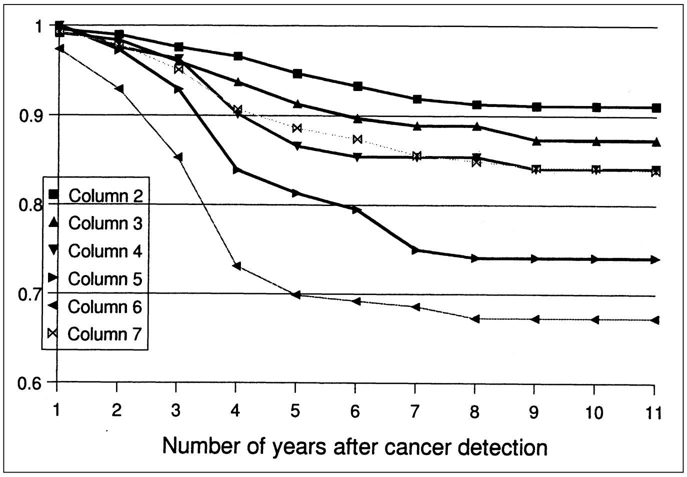


Fig. 1 Survival curves for the covariable *Number of positive nodes.* Column 2 is “No invaded ganglion”; Column 3 is “1 invaded ganglion”; Column 4 is “2 invaded ganglions”; Column 5 is “3 to 4 invaded ganglions”; Column 6 “More than 4 invaded ganglions”; Column 7 is “Mean survival”.

5. Application to Breast Cancer Survival

To illustrate the theory presented above, we used data from an investigation carried out in Marseille by Professor J.-M. Spitalier and Doctor D. Hans. In their study, 1304 patients were observed during 10 years and 45 variables were recorded for each patient. A treatment was applied to the different patients according to the values of the variables.

Fifteen out of 45 variables were retained after a preliminary data analysis: 1) age, 2) clinic class, 3) thermographic class, 4) senographic class, 5) echographic class, 6) clinic PEV, 7) clinic diameter, 8) clinic behavior, 9) side, 10) histology, 11) N histology, 12) number of positive nodes (number of metastatic lymph nodes), 13) estradiol receivers, 14) progesterone receivers, 15) UICC stage.

Computations were programmed in APL. The first step allowed to select 5 covariables which were: the thermographic class, the number of positive nodes, the histology, the N histology and the UICC stage using the following thresholds: p_1 = 5% and α = 5% *(cf. Fig. 3)*.

The second step confirmed the choice, in the sense that combinations of these covariables were selected, with the following thresholds: p_2 = 15% and α = 5% *(cf. Fig. 4)*. Nevertheless, at the third step, the best combination of three covariables was the age, the histology and the number of positive nodes. That is, in this step age was selected while this covariable was not selected earlier. This combination gave about 40% of relative information on the survival time.

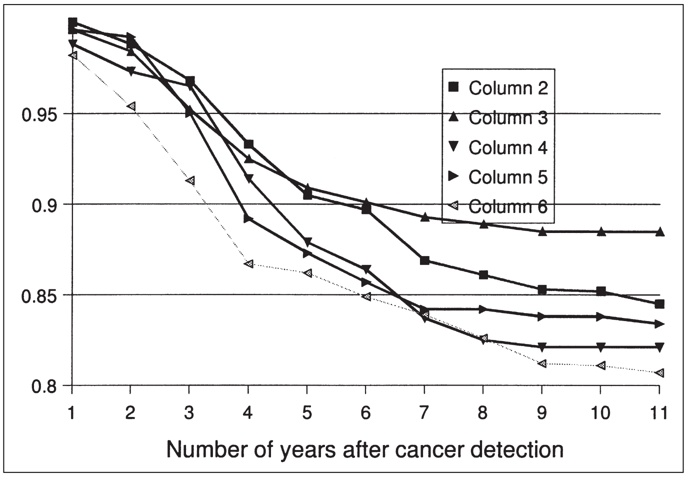


Fig. 2 Survival curves for the covariable *Age.* Column 2 is “25 to 44 years”; Column 3 is “45 to 51 years”; Column 4 is “52 to 59 years”; Column 5 is “60 to 69 years”; Column 6 is “More than 70 years”.

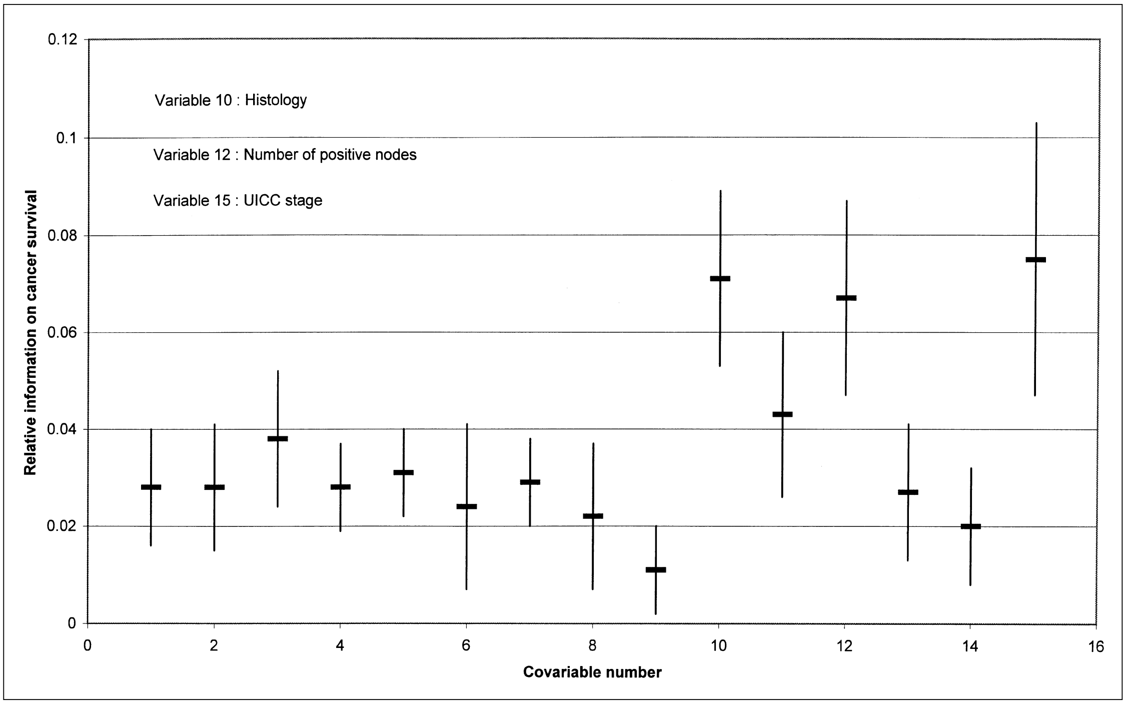


Fig. 3 Relative information confidence interval for each covariable.

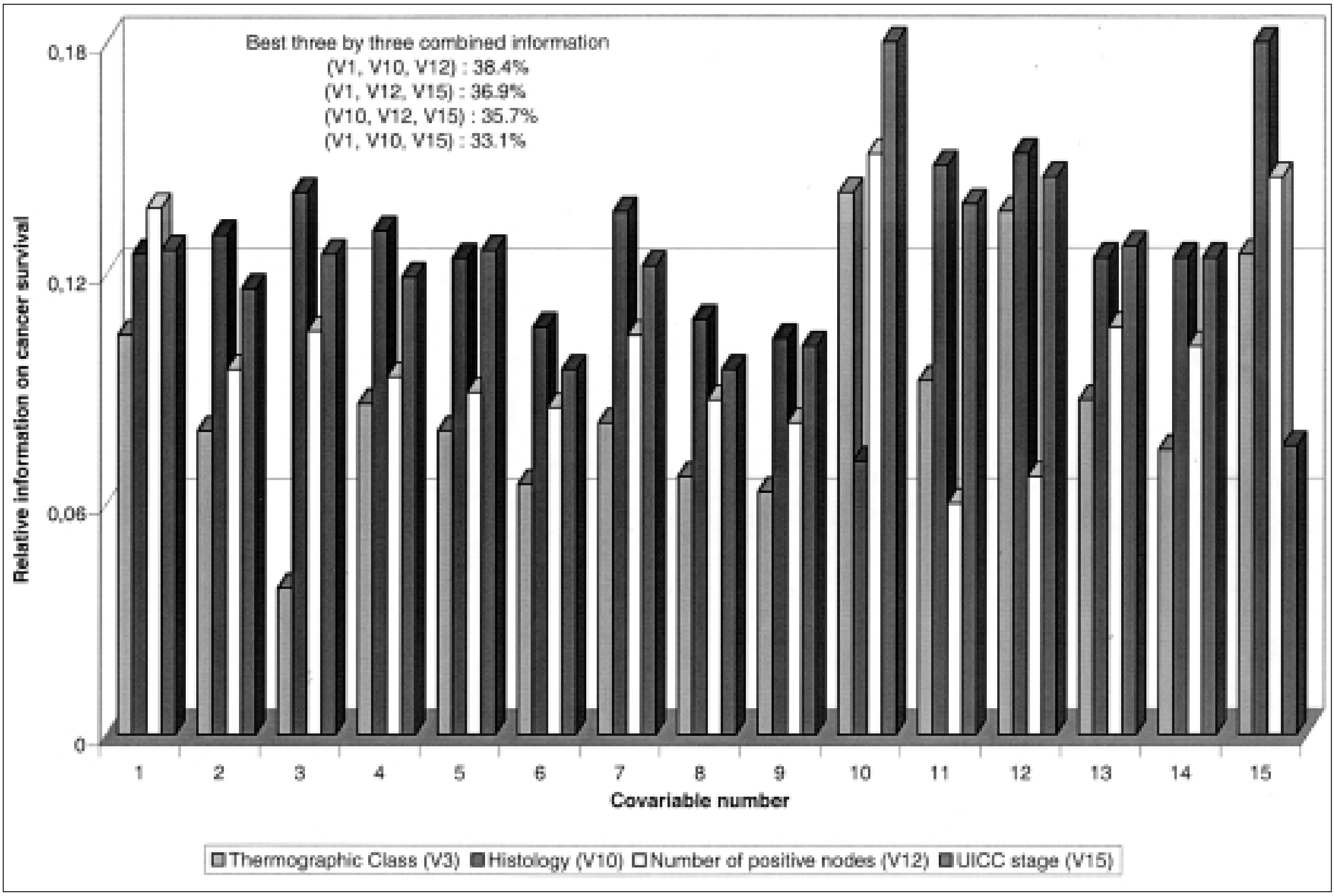
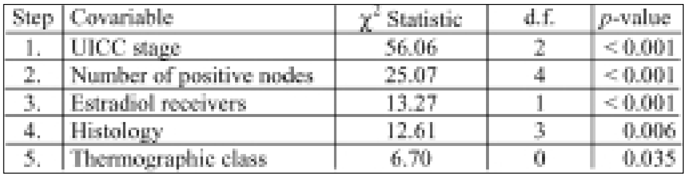


Fig. 4 Relative information for two by two combinations between the 15 covariables and respectively V3, V10, V12, V15. V1 is the covariable *Age.*

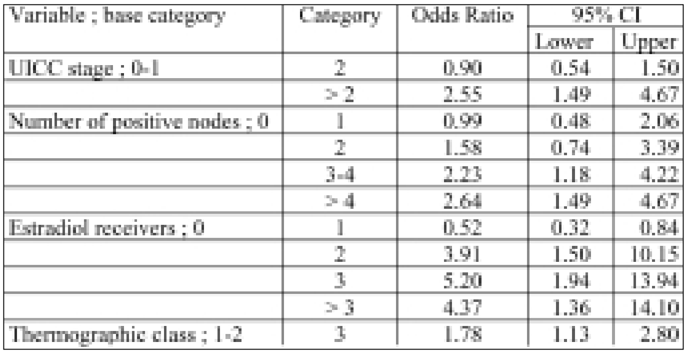
When a covariable has a high prognostic value, it is interesting to be able to assign a survival curve to the different possible values of this covariable *(cf. Figs. 1 and 2).* These curves represent the survival rate conditional to the value taken by this covariable on a patient. They are useful tools for medical prognostics.

The studies which were undertaken separately from the four covariables (age, histology, number of positive nodes and UICC stage) indicate the modalities which have the best survival expectation. So, for the covariable number of positive nodes *(cf. Fig. 1)*, 0, 1 or 2 positive nodes give survival curves that are near and above the average survival whereas the modality "3 to 4 positive nodes" is clearly less than average. The modality "more than 4 positive nodes" is the less favorable. If the covariable age is taken individually, it carries little information on the survival variable: this is shown by the entwining of the survival curves associated with its modalities *(cf. Fig. 2).*



Step	Covariable	χ^2 Statistic	d.f.	p-value
1.	UICC stage	56.06	2	< 0.001
2.	Number of positive nodes	25.07	4	< 0.001
3.	Estradiol receivers	13.27	1	< 0.001
4.	Histology	12.61	3	0.006
5.	Thermographic class	6.70	0	0.035

Table 1 Stepwise (forward) building up of the logistic model.



Variable ; base category	Category	Odds Ratio	95% CI	
			Lower	Upper
UICC stage ; 0-1	2	0.90	0.54	1.50
	> 2	2.55	1.49	4.67
Number of positive nodes ; 0	1	0.99	0.48	2.06
	2	1.58	0.74	3.39
	3-4	2.23	1.18	4.22
	> 4	2.64	1.49	4.67
Estradiol receivers ; 0	1	0.52	0.32	0.84
	2	3.91	1.50	10.15
	3	5.20	1.94	13.94
	> 3	4.37	1.36	14.10
Thermographic class ; 1-2	3	1.78	1.13	2.80

Table 2 Final results for the logistic regression after 5 steps.

6. Comparison with other Selection Methods

We compared the selection method presented above with analogous ones based on the binary logistic regression model and the semiparametric Cox's regression model [1] which are very popular methods in survival analysis. All calculations were carried out with BMDP version PC90. It is worth noticing that, in practice, the true model is unknown so that a performance comparison of different predictor selection methods can be carried out only through extensive model simulations. This problem has not been addressed here and will be presented in a subsequent paper.

6.1 Binary Logistic Regression

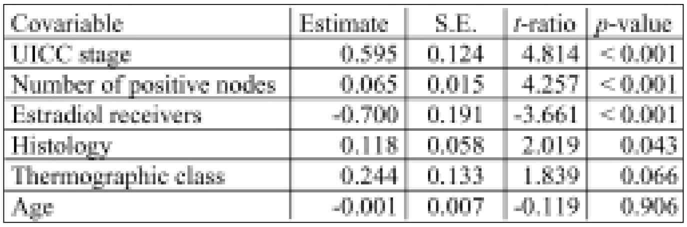
We have applied the binary logistic regression (dead within 11 years, alive after 11 years) to the data with the 15 covariables presented in Section 5. We have chosen the forward stepping procedure, which means that at the beginning none of the covariables are in the model. The algorithm stops when the p-value for the improvement chi-squared statistics is greater than 0.1. The results of the stepwise procedure are presented in Table 1. We can see that 4 out of the 5 selected covariables are also selected by our method. Table 2 shows the influence of the selected covariables on the survival. As for our method, the number of positive nodes significantly decreases the probability of survival. On the other hand, the covariable age does not improve the fit of the logit model at all.

6.2 Cox's Regression Model

We have also used Cox's regression model to estimate the influence of the four covariables selected by our nonparametric method and two others selected by stepwise procedure in the logistic regression. Thus, the suggested model contains six unknown parameters of interest, each of them associated with these six covariables. The results of the estimation procedure are presented in Table 3. They are similar to our results since the UICC stage, the number of positive nodes and the histology are closely related to the survival risk of Cox's model. The regression parameter for age is not significantly different from zero in this model.

7. Conclusions

We have presented an asymptotic and nonparametric method for predictor selection which can be applied when the number of statistical units is large and when there is no reasonable model for the observed phenomena. Using this new entropy-based stepping procedure, some covariables carrying most of the information on the variable of interest Y can be selected. The correlation between Y and the covariables is quantified by the relative information of this set of covariables on Y, that is the proportion of reduction of uncertainty on Y by knowing the covariables. When this variable is the survival time after a disease detection, survival curves for different modalities of a selected predictor can be aggregated using an entropy conservation principle which



Covariable	Estimate	S.E.	t-ratio	p-value
UICC stage	0.595	0.124	4.814	< 0.001
Number of positive nodes	0.065	0.015	4.257	< 0.001
Estradiol receivers	-0.700	0.191	-3.661	< 0.001
Histology	0.118	0.058	2.019	0.043
Thermographic class	0.244	0.133	1.839	0.066
Age	-0.001	0.007	-0.119	0.906

Table 3 Cox regression parameter estimation.

implies no significant information loss after aggregation. The survival prediction is then simplified.

When the proposed method was applied to patients having breast cancer, four variables were selected. A comparison with predictor selection methods based on the binary logistic regression and Cox's regression gave comparable results. The best predictors obtained are the number of positive nodes, the UICC stage, the histology and the thermographic class. Only our method selected age in the set of significant predictors.

We focused here on the presentation of the theory and its illustration. Nevertheless, a performance study of this selection procedure with other methods is under preparation through extensive model simulations. The first results obtained show good performance when the number of observation units is large and our method compares favourably with model-based methods when the model is mispecified (i.e., the known simulated model is).

REFERENCES

1. Cox DR. Regression models and life-tables (with discussion). J Roy Statist Soc B 1972; 34: 187-220.
2. Daley DJ, Vere-Jones D. An introduction to the theory of point processes. New York: Springer 1988.
3. Darbellay GA. A nonparametric estimator for the mutual information. Proceedings of Stochastics '98, Prague 1998; 93-8.
4. El Hasnaoui A. Le concept du gain d'information: une nouvelle approche en épidémiologie quantitative. Thèse de doct. – Univ. de Montpellier 1993.
5. Kempton RA. The structure of species abundance and measurement of diversity. Biometrics 1979; 35: 307-21.
6. Kent JT. Information gain and a general measure of correlation. Biometrika 1983; 70: 163-74.
7. Kent JT, O'Quigley J. Measure of dependence for censored survival data. Biometrika 1988; 75: 525-34.
8. Kullback S. Information theory and statistics. New York: Wiley 1959.
9. Manuceau J, Troupe M, Vaillant J. Information and prognostic value of some variables on breast cancer. ESAIM 2000; in press.
10. Manuceau J, Troupe M, Vaillant J. On an entropy conservation principle. J Appl Prob 1999; 36: 607-10.
11. Mathai AM, Rathie PN. Characterization of Matusita's measure of affinity. Ann Inst Statist Math 1972; 24: 473-82.
12. Nayak TK. Sampling distributions in analysis of diversity. Sankhya series B 1986; 48: 1-9.
13. Rao CR. Diversity: Its measurement, decomposition, apportionment and analysis. Sankhya Series A 1982; 44: 1-22.
14. Rao CR. Generalization of ANOVA through entropy and cross entropy functions. In: Probability theory and mathematical statistics (vol 2). Utrecht: Sciences Press 1986; 477-94.
15. Rényi A. On measures of entropy and information. Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability 1961; 1: 547-61.
16. Robert C. An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. J Appl Prob 1990; 27: 303-13.
17. Shannon, CE. A mathematical theory of communication. Bell Syst Tech J 1948; 27: 379-423 and 623-56.

Address of the authors:
Dr. J. Vaillant,
Université des Antilles et de la Guyane,
UFR Sciences Exactes et Naturelles,
Campus Fouillole; 97159 Pointe-á-Pitre
Guadeloupe, F.W.I.
E-mail: Jean.Vaillant@univ-ag.fr