

Research Article

Classification and Regression Trees on Aggregate Data Modeling: An Application in Acute Myocardial Infarction

**C. Quantin,^{1,2} L. Billard,³ M. Touati,⁴ N. Andreu,^{1,2} Y. Cottin,⁵
M. Zeller,⁵ F. Afonso,⁴ G. Battaglia,⁴ D. Seck,⁴ G. Le Teuff,^{1,2}
and E. Diday⁴**

¹ INSERM EMI 0106, 21000 Dijon, France

² Université de Bourgogne, Service de Biostatistique et Informatique Médicale, CHU, 21000 Dijon, boulevard Jeanne d'Arc BP 77908, 21079 Dijon Cedex, France

³ Department of Statistics, University of Georgia, Athens, GA 30602-1952, USA

⁴ CEREMADE CNRS UMR 7534, Université de Paris, Dauphine 75775 Paris Cedex 16, France

⁵ Service de Cardiologie, Centre Hospitalier du Bocage, BP 77908, 21079 Dijon Cedex, France

Correspondence should be addressed to C. Quantin, catherine.quantin@chu-dijon.fr

Received 22 October 2010; Revised 24 March 2011; Accepted 25 May 2011

Academic Editor: Peter van der Heijden

Copyright © 2011 C. Quantin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cardiologists are interested in determining whether the type of hospital pathway followed by a patient is predictive of survival. The study objective was to determine whether accounting for hospital pathways in the selection of prognostic factors of one-year survival after acute myocardial infarction (AMI) provided a more informative analysis than that obtained by the use of a standard regression tree analysis (CART method). Information on AMI was collected for 1095 hospitalized patients over an 18-month period. The construction of pathways followed by patients produced symbolic-valued observations requiring a symbolic regression tree analysis. This analysis was compared with the standard CART analysis using patients as statistical units described by standard data selected TIMI score as the primary predictor variable. For the 1011 (84, resp.) patients with a lower (higher) TIMI score, the pathway variable did not appear as a diagnostic variable until the third (second) stage of the tree construction. For an ecological analysis, again TIMI score was the first predictor variable. However, in a symbolic regression tree analysis using hospital pathways as statistical units, the type of pathway followed was the key predictor variable, showing in particular that pathways involving early admission to cardiology units produced high one-year survival rates.

1. Introduction

Predictive variables of in-hospital mortality after acute myocardial infarction (AMI) have been studied for many years and include age [1], the severity of disease at admission [2], previous history of AMI, family history of coronary disease [3], and comorbidities such as excess weight [4, 5], diabetes [6–8], systemic hypertension [9], and dyslipidemia [10]. As part of these clinical variables, the time between occurrence of the first symptoms and admission into intensive care units (ICU) or into cardiology units represents another very well-known determinant of in-hospital mortality after myocardial infarction (MI), as this is a measure of the quickness of therapeutic intervention and care [11]. It is well known that the severity of the AMI at admission governs the process of care especially the decision(s) to move the patient to an intensive care unit or to cardiology. Earlier studies such as [12–14] have suggested that patients treated by cardiologists had better survival rates than those treated by noncardiologists. In [15], rather than treatment by a cardiologist, attention focused on admission to a cardiology unit.

It is, therefore, natural that cardiologists and other clinical researchers in general are becoming increasingly interested in the role and impact of the pathway followed by a patient as a potential factor associated with survival, taking into account the influence of other prognostic factors. Hospital pathway herein is defined as the temporal process followed covering all hospital stays within the same hospital and across multiple healthcare institutions. A simple pathway is one where a patient is admitted to, for example, a cardiology unit and then discharged or a more complicated pathway is when the patient after admission to an intensive care unit (ICU) is subsequently moved to a cardiology unit at the same or another hospital. There are many possible pathways involving patients moving in and out of cardiology and noncardiology units not necessarily at the same institution.

As a consequence, the aim is not only to consider the effect of a patient's admission directly into a cardiology unit compared to admission into an unspecialized hospital unit as was done in [15] but also to take into account the complexity of the patient's pathways for treatment of AMI across multiple hospitals and multiple units. We have two questions: how can complex pathways be considered in the analysis, and what is the influence of these pathways on prognosis?

Unfortunately, standard analyses in which pathway type is a covariate in the model are questionable, because the pathway itself may not only affect the outcome but also depend on some patients' characteristics associated with survival. Therefore, pathway type is partly a mediating variable for other prognostic factors. If so, then adjusting these other factors for pathway, and vice versa, is problematic and may bias the results.

To avoid such problems, an alternative approach is to replace the "individual" data analysis by an ecological comparison of "groups of patients", with patients following the same pathway forming a group. In fact, this method takes into account the hierarchical structure of the data (e.g., the patients hospitalized in the same hospitals and following the same pathways may share the same sociodemographics characteristics such as place of residence and clinical characteristics such as the severity of the disease). Thus, rather than attempting to estimate the effects of these other factors adjusted for pathway, the ecological analysis would investigate which characteristics of patients who follow the same pathways are associated with better survival rates.

This poses two problems. First, it is often difficult to analyse a large number of modalities especially for multilevel analysis. Second, these analyses of pathways requires aggregation of those patients who follow a given pathway. Standard ecological analyses of

these aggregated data replace individual data values by their group-level aggregate values, usually the mean value for continuous variables or proportional values for binary variables. These values reflect central tendency only, ignoring variations among those individuals who made up that group or pathway. Yet, ignoring this variation in the predictor values across individuals in the same group is known to induce systematic bias; see, for example, [16–19].

So-called symbolic analytic methods provide an alternative approach to analyzing grouped data. In contrast to standard ecological analyses, symbolic analyses take into account not only the central tendency but also the internal variations of the individuals within the grouped data. This is achieved by retaining all the aggregated information as, for example, lists, interval- or histogram-valued data, instead of replacing the individual values by the one central tendency value; see Section 2.3 for a description of symbolic data. It is possible to show that the total variation of (say) interval-valued data equals the sum of the within intervals variations and the between intervals variation [20]. Thus, ecological analyses which use the central tendency are ignoring the within observation variations and so base their results on the between observations (i.e., groups) variations only. This loss of information clearly impacts on the results of the analyses. Attempts have been made to overcome this by replacing the observed interval value by its end points values; the results are unsatisfactory and tend to reflect those obtained when using central tendency values. In both cases, the internal variations in the data are not observed by, nor used in, an ecological or standard analysis.

The objective of the present study is to illustrate potentially new insights that could be obtained by a symbolic analysis of the role of hospital pathways on the survival of patients one year after acute myocardial infarction. To this end, we compare results of a conventional classical CART (classification and regression trees, [21]) analysis on individual patients and on the ecological pathways central tendency values, and on the symbolic values of the pathways using a symbolic version of the classical CART methodology. One feature of this symbolic CART methodology is that when observations are classical values, the classical CART results come out as a special case of the symbolic CART. The methods are described briefly in Section 3 with the results in Section 4.

2. Data

2.1. The Patients

The study population was made up of 1095 patients who were hospitalized for myocardial infarction (AMI) at 6 healthcare institutions: the teaching hospital of Dijon, one private hospital located in the suburbs of Dijon (Fontaines), and 4 public local hospitals (Beaune, Châtillon, Montbard, and Semur) over a period of 18 months (from June, 2001 to December, 2002) in the French administrative area of the Côte d'Or. One hospital had three different cardiology units, while the others had one such unit, to which patients could be admitted. These patients were identified through the French Registry of AMI of the Côte d'Or, where all AMI cases hospitalized in public or private hospitals in this area are recorded.

Patient characteristics were drawn from this French registry which also contains information on treatment and patient outcomes. These clinical data were linked to the health administrative database in each hospital in order to collect information on hospital stay using the French Diagnosis Related Groups (DRGs) information system [22]. This system describes hospital activity by recording a discharge abstract for each patient stay.

Table 1: Random variables.

Y_j	(a) Variable Description	Categories \mathcal{O}_j	(b) A pathway observation
Y_1	Gender	{Female, Male}	{Female, 0.1; Male, 0.9}
Y_2	Angina	{Yes, No}	{No, 1.0}
Y_3	Smoker	{Yes, No}	{Yes, 0.9; No, 0.1}
Y_4	Heredity	{Yes, No}	{Yes, 0.4; No, 0.6}
Y_5	Prior AMI	{Yes, No}	{No, 1.0}
Y_6	Diabetes	{Yes, No}	{Yes, 0.1; No, 0.9}
Y_7	Arterial hypertension	{Yes, No}	{Yes, 0.9; No, 0.1}
Y_8	Cholesterol	{Yes, No}	{Yes, 0.1; No, 0.9}
Y_9	ST deviation	{Yes, No}	{Yes, 0.1; No, 0.9}
Y_{10}	TIMI score	{0, 1, ..., 14}	{1, 0.11; 2, 0.11; 3, 0.11; 4, 0.22; 5, 0.11; 7, 0.33}
Y_{11}	Expert class	{1, ..., 8}	{7, 0.5; 8, 0.5}
Y_{12}^a	Survived one year	{Yes, No}	{Yes, 0.4; No, 0.6}

^a Explanatory variable.

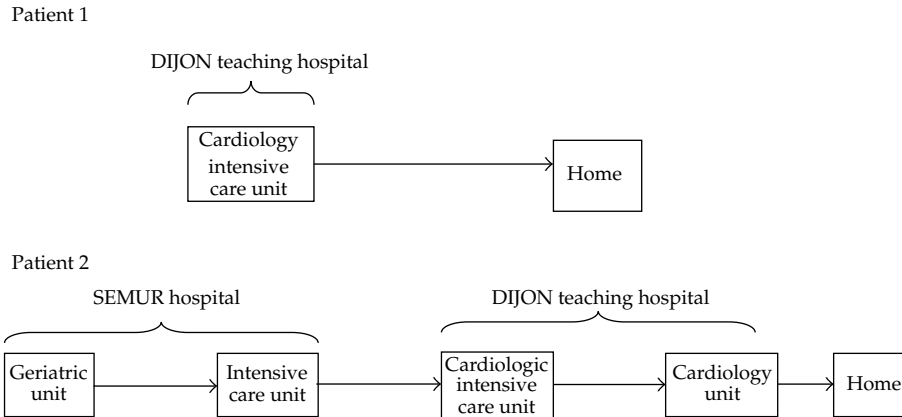
2.2. Description of Variables

From the 218 measurements recorded per patient, eleven variables were selected as being important for the present study. These were gender (gender: female, male); did the patient have a previous experience of acute myocardial infarction or angina (angina: yes, no); did s/he have a history of smoking (smoker: yes, no); is there a family history of coronary disease, of myocardial infarction or angina (heredity: yes, no); had the patient experienced a prior myocardial infarction (prior AMI: yes, no); was the patient diabetic (diabetes: yes, no); was the patient being treated for systemic arterial hypertension (arterial hypertension: yes, no); did the patient have dyslipidemia and treated for hypercholesterol (cholesterol treated: yes, no); did the patient experience acute myocardial infarction with or without ST-segment deviation (ST elevation: yes, no); severity of disease at admission, that is, thrombolysis in myocardial infarction (TIMI) score (TIMI: 0, ..., 14). The TIMI risk score is a weighted integer score based on 8 clinical risk indicators that can be easily ascertained at presentation (see Table 1 in [23]). For each patient, the score is calculated as the arithmetic sum of the points for each risk feature present. Note that obesity per se was not considered as a separate variable as measures of excess weight are factored into the TIMI score likewise age.

The eleventh variable, called ‘‘Expert Class’’ (Expert: 1, ..., 8) related to the type of healthcare at admission; see Section 2.3. It is this concept that is the focus of the present study. Also selected as the explanatory variable in the regression tree analysis was whether or not the patient survived at least one year after admission. A summary of these variates is provided in Table 1(a).

2.3. Hospital Pathways

The patient hospital pathway was defined as the temporal process covering all hospital stays within the same hospital and across multiple healthcare institutions by using the discharge abstracts; that is, it is the exact order that a patient followed. For example, a patient may be hospitalized directly in one of the 8 intensive care units (ICU) of cardiology of the Côte d’Or. However, some patients may be hospitalized in another medical unit, such as geriatrics



Hospital pathway for treatment of first acute myocardial infarction. (Hospitalizations in 2 health care centers consecutive)

Figure 1: Two examples of pathways.

or neurology before being cared for in a cardiology unit, in so far as AMI diagnosis is not systematically assessed at the patient's admission. Moreover, transfers between hospitals may be needed in order to perform complementary investigations such as revascularization.

Examples of two possible pathways are depicted in Figure 1. Patient 1 was admitted directly into a cardiology intensive care unit at the Dijon Teaching Hospital before being released (and sent home). Patient 2 was admitted to the geriatric unit at the Semur Hospital and moved to the Semur intensive care unit; this patient was then transferred to the Dijon Teaching Hospital first to the cardiology intensive care unit and later to the cardiology unit and was subsequently released and sent home.

To identify the different types of pathways, we started by ordering the first hospital unit according to their frequencies by which admissions occurred. Then, for each of those hospital units, we added the second hospital unit to which a patient was transferred, and we ordered all the resulting paths by their frequencies and so on (with the triples, quadruples, etc.) as long as the "tuple" frequencies remained higher than a given threshold. The pathways that emerged were called the "strong pattern" pathways. These are described in Table 2. Also shown are the number of patients who followed each type of pathway, as well as, the number of different pathways matching that pattern description. A hospital pathway is based on three types of information: the type of healthcare institution, the type of hospital unit, and the chronological order of admission to the unit.

Each patient followed a pathway. There were $m = 204$ distinct pathways for the 1095 patients in the study. These m pathways fell into one of ten pathway patterns; see Table 2. Thus, pattern 1 identifies 8 of the 204 pathways as being such that a patient was admitted to a cardiology unit at a single hospital before being discharged and sent home; for example, a cardiology unit at Dijon Hospital (as in Patient 1 in Figure 1). This pattern accounted for 403 of the patients. On the other hand, if a patient spent time at more than one institution and was admitted to noncardiology units before eventual admission to cardiology (e.g., Patient 2 of Figure 1), then the pattern 10 pertains. The aggregation of individual patient values by pathway performs symbolic data.

Table 2: “Patterns” of hospital pathways.

Pattern	Expert	Description	Pathways		Patients	
	Class		#	%	#	%
1	1	Single unit: UF1*: cardiology	8	3.9	403	36.8
2	2	Single unit: UF1: no cardiology	2	1.0	2	0.2
3	3	Transfer: single hospital, multiunits UF1: cardiology, UF2**: no cardiology	10	4.9	14	1.3
4	3	Transfer: single hospital, multiunits UF1: cardiology, UF2: cardiology	48	23.5	427	39.0
5	4	Transfer: single hospital, multiunits UF1: no cardiologist, UF2: no cardiology	5	2.5	5	0.4
6	5	Transfer: single hospital, multiunits UF1: no cardiology, UF2: cardiology	32	15.7	43	3.9
7	6	Transfer: multi hospital, multiunits UF1: cardiology, UF2: no cardiology	2	1.0	3	0.3
8	7	Transfer: multi hospital, multiunits UF1: cardiology, UF2: cardiology	78	38.2	177	16.2
9	8	Transfer: multi hospital, multiunits UF1: no cardiology, UF2: no cardiology	1	0.5	1	0.1
10	8	Transfer: multi hospital, multiunits UF1: no cardiology, UF2: cardiology	18	8.8	20	1.8

* UF1: first hospital unit **UF2: second hospital unit.

Since this pathways variable only takes into account the type of hospital unit and the chronological order of the patient’s transfer, the type of healthcare institution at admission was introduced as another covariate (called Expert Class) in the analysis.

2.4. Symbolic Data

Symbolic data methodology extends standard classical categorical or numerical variables to the case where the variable values may be intervals, histograms, lists of categorical values, and so forth. These kinds of variable values emerge when we aggregate observations measuring the clinical information between patients within the same pathways. To illustrate, consider the observations of Table 3, where for each patient the first hospital unit to which they were admitted is identified, and suppose also that age is recorded by year and smoking history is recorded as one of no, light, heavy smoker. When the focus of interest is the admitting hospital rather than the individual patients per se, the statistical unit is now that hospital. Therefore, when we aggregate over those patients who were admitted to that hospital, we obtain (for the classical data of Table 3), the symbolic data of Table 4. Thus, the variable age has the observed interval value of [70, 82] for Hospital 1 and [69, 80] for Hospital 2. Classical values are special cases; for example, for Hospital 3, age is in the interval [76,76]. The variable for smoking history has a realization for Hospital 1 that reads as 25% were light and 75% were heavy smokers. This variable is now a modal-valued categorical variable. Categorical symbolic realizations are lists of possible (discrete) values. When no weights are attached, it is assumed the listed outcomes are equally likely. Interval-valued realizations are assumed to take specific values that are uniformly distributed across that interval.

Table 3: Sample of patient records.

Patient	Hospital	Age	Smoker
Patient1	Hospital 1	74	heavy
Patient2	Hospital 1	78	light
Patient3	Hospital 2	69	no
Patient4	Hospital 2	73	heavy
Patient5	Hospital 2	80	light
Patient6	Hospital 1	70	heavy
Patient7	Hospital 1	82	heavy
Patient8	Hospital 3	76	no
⋮	⋮	⋮	⋮

Table 4: Symbolic data by hospital.

Hospital	Age	Smoker
Hospital 1	[70, 82]	{light 1/4, heavy 3/4}
Hospital 2	[69, 80]	{no, light, heavy}
Hospital 3	[76, 76]	{no}
⋮	⋮	⋮

Sometimes, because of underlying conditions, it can be that what seem to be rectangular observations are more accurately hypercubes. For example, suppose the two random variables Y_1 and Y_2 take values on the intervals $[100, 120]$ and $[95, 115]$, respectively. Then, the observed value is the rectangle $\text{Rect} = [100, 120] \times [95, 115]$. Suppose, however, that $Y_1 \geq Y_2$ (as, e.g., when systolic blood pressure is greater than diastolic blood pressure). Under this condition, values inside the triangle with vertices $(100, 100)$, $(100, 115)$ and $(115, 115)$ are not possible. Thus the effective observation space is the hypercube with vertices $(100, 95)$, $(120, 95)$, $(120, 115)$, $(115, 115)$, and $(100, 100)$.

In general, datasets become symbolic-valued after aggregation of a larger dataset of classical or symbolic valued observations. Many forms of aggregation could apply for any one dataset, the actual aggregation(s) adopted being driven by the scientific questions behind the study itself. For example, an insurance company (with its database running into millions of individual records) is less interested in one specific person's medical-care usage than it is interested in the pattern of forty-year old women (say) or sixty-year old cardiac victims, and so on.

To use standard methodology on the resulting symbolic data will be inadequate. For example, suppose that a person's weight W fluctuates over the interval $W_1 = [160, 166]$ in pounds and a second person's weight ranges across $W_2 = [157, 169]$. When we take the classical valued midpoint (or, more generally, a central tendency measure), here 163 in each of W_1 and W_2 , we would obtain the same results. Yet, these observations are differently valued and any analysis should reflect these differences. Thus, an analysis on the midpoints ignores the information that the first interval W_1 is internally less variable than is the second W_2 (with internal variations of 3 and 12, resp.). Symbolic analytic techniques take these internal variations into account.

In a different direction, symbolic data can be used to maintain data confidentiality. For example, rather than pulse rate = 64, the value can be recorded as $64 - \delta_1, 64 + \delta_2$ (with $\delta_1 \neq \delta_2$),

for example, pulse rate = [60,71]. In the current application, we replaced the description of each of the patients by the description of pathways; thus, it is difficult to identify patients within a given pathway.

Some measurements are inherently imprecise or fuzzy. Rather than trying to take a single measure such as the mean to serve as the surrogate for the measurements, the range of actual values over a specific time period can be retained as an interval-valued datapoint (or, histogram-valued, or other suitable symbolic-valued variable). For example, [24] (along with several other researchers such as [25]) discusses how the level of systolic pressure and its changes over time are an important aspect on strokes and how difficult it is to obtain an accurate single measurement due to its highly variable nature especially over time. The study [24] goes on to describe how the researcher can be misled if the left ventricular hypertrophy is used as a marker for systolic blood pressure. In contrast, if, for example, all systolic measurements over the time period were aggregated, then a symbolic valued observation emerges. In some cases (e.g., smoking levels, other social and educational measures; see, e.g., [26]), issues surrounding residual confounding (but not necessarily confounding in and of itself) are obviated (see, e.g., [27]).

In the context of the current study, the scientific question revolves around the pathways (or patterns) encountered by cardiac patients rather than any one patient *per se*. Are certain pathways more or less likely to predict survival; are certain predictor variables more indicative of particular pathway tracks, and so on? The example of Tables 3 and 4 would represent a pathway of but one step, the admitting hospital. In the actual study, the pathways are those described earlier; see Table 2. The observed realization for the outcome and predictor variables is what pertains when the observed values for the individual patients, who collectively make up a pathway, are aggregated. These resulting observations are perforce symbolic valued being lists, interval-valued, and/or modal-valued, as appropriate. Table 1(b) gives the results for one such pathway. For a more extensive review of symbolic data, see [28–30].

3. Methods

In order to select the potential covariates associated with one-year survival, we use the CART method [21]. This nonparametric approach is interesting when the type or nature of candidate covariates is important. How this applies in the standard classical setting when the units are individual patients is described in Section 3.1. When the units are the pathways, there are two methods. First, the ecological pathway values can be used as described in Section 3.2. The adaptation of the classical CART method to the pathways symbolic data is discussed in Section 3.3.

3.1. Regression Tree Analysis (Classical CART Method) on Patients

To explore the potential determinants of one-year mortality after AMI, we first used the usual classification and regression tree (CART) method [21] on the 1095 individual patients themselves (so as to compare these results with the counterpart results based on the pathways). A CART analysis uses regression techniques in order to find a classification that best explains the relationships between given categorical variables. The predictor/regression variables (such as weight, TIMI score, and smoker, etc.) are used to inform the analyst how the classification tree should be constructed. This is now a well-established technique and

has been applied extensively to a wide variety of statistical settings. Recent use of CART to cardiology patients in a hospital environment can be found in, for example, [21, 22, 31], and in cancer [32].

3.2. Regression Tree Analysis (Ecological CART Method) on Pathways

The primary focus, however, is on pathways as the statistical unit. Note that any particular pathway has one or more individual patients who follow the same pathway. For an ecological analysis, central tendency values of the individual values in each pathway are used as the “observation” for each variable. Then, the standard classical CART methodology as described above is run using these “observation” values. For our study, the mean values were calculated, and hence, the CART methodology was applied to the 204 pathways.

3.3. Regression Tree Analysis (Symbolic CART Method) on Pathways

Rather than taking a central tendency value as in the ecological analysis, the pathway values were retained as symbolic observations such as lists or intervals; see Section 2.3. Since an “observation” for a pathway will in general be a hypercube in p -dimensional space rather than the point in p -dimensional space of standard/classical data associated with a patient, a so-called symbolic CART analysis is required. This is achieved by combining the concepts of symbolic regression analysis (see, e.g., [33–35]) with the basic ideas of the CART method [21] as follows.

Both the classical CART method and the symbolic regression tree analysis are based on a divisive top-down classification tree method, restricted to recursive binary partitions, until a suitable stopping rule prevents further divisions. The binary partition is induced by the variable which implies the best binary splitting of the variables for a given criterion. The criterion used by the CART method is a discrimination criterion $D(N)$ which measures the impurity of a node N with respect to the prior partition (G_1, \dots, G_J) . Like the classical CART, the discrimination criterion used for the symbolic CART is the Gini measure. The Gini measure of node impurity is a measure which reaches a value of zero when only one class is present at a node. With priors estimated from class sizes, the Gini measure is computed as the sum of products of all pairs of class proportions for classes present at the node; it reaches its maximum value when class sizes at the node are equal. Therefore, following Breiman et al. [21],

$$D(N) = \sum_{i \neq f} p_i p_f = 1 - \sum_{i=1, \dots, J} p_i^2, \quad (3.1)$$

with $p_i = n_i/n$, $n_i = \text{card}(N \cap G_i)$ and $n = \text{card}(N)$ in the classical case. In the symbolic case, $n_i =$ number of individuals belonging to N which verify the current description of N and at the same time belong to G_i and $n =$ the total number of the individuals belonging to N . To normalize $D(N)$, we multiply by $J/(J - 1)$ where J is the number of prior classes; it then lies in the interval $[0, 1]$.


To construct the tree, we partition the node N into two nodes $N = (N_1, N_2)$. The idea is to find that node and partitioning which maximizes the reduction in total variation. That is, we want to maximize

$$\Delta D = D(N) - D(N_1, N_2) = D(N) - D(N_1) - D(N_2). \quad (3.2)$$

If the variable Y_j is the partitioning variable from the complete set of variables $\mathbf{Y} = (Y_1, \dots, Y_p)$, then the cut point is c_j , say. For modal multivalued data (as in our dataset), we find this c_j as follows. Suppose that the set of possible categorical values for each Y_j is $\mathcal{O}_j = \{m_{j1}, \dots, m_{js}\}$, $j = 1, \dots, p$. Then, a particular observation u takes the values $Y_{uj} = \{m_{ujk}, p_{ujk}; k = 1, \dots, s\}$ for $u = 1, \dots, n_i$, $j = 1, \dots, p$, where n_i is the number of observations at node N and p_{ujk} is the probability (or relative frequency) that the k th category m_{jk} occurred for the j th variable by the u th observation, with $\sum_k p_{ujk} = 1$. (When a particular category does not occur, its probability is zero.) For each k in turn, the p_{ujk} are ordered from smallest to largest. Then, the cut point, c_{jkr} , for the category m_{jk} is the probability

$$c_{jkr} = \frac{(p_{jkr} + p_{jk,r+1})}{2}, \quad r = 1, \dots, n_k - 1, \quad k = 1, \dots, s, \quad (3.3)$$

where $n_k < s$ is the number of distinct values of p_{jkr} . Likewise, pairs of categories m_{ujk_1} and m_{ujk_2} have probability $(p_{ujk_1} + p_{ujk_2})$ which are ordered from the smallest to the largest, and then the potential cut point between the pairs of categories is the mean of these probability pairs. Similarly, in sets of three, four, and so on, categories from \mathcal{O}_j are taken and ordered. The methodology then considers all possible partitions based on these possible cut points to find that partition which maximizes ΔD of (3.2).

The methodology is performed by the STREE (symbolic regression TREE) algorithm.  The algorithm includes boosting, bagging, and cross-validation components. A simple example illustrating the method is given in [36]. In [36], a comparison of the classical CART and STREE on learning and test datasets (in a variety of scenarios) showed that the misclassification rate for STREE was considerably less than for CART. Note that when all the input variables have classical point values (e.g., classical $x = a$ is equivalent to symbolic $x = [a, a]$), the STREE methodology is the same as the CART methodology.

Here, we apply the regression tree method to the whole population (considered as a class) in two ways. First we refer to the classical regression tree as the case where the units are the patients described by their (classical) numerical or categorical clinical variables (as developed by the CART algorithm). In the second case, we apply the symbolic regression tree methodology where the units are the pathways described by symbolic-valued variables.

4. Results

4.1. Description of Pathways

The mining of the administrative database for the frequent sequential patterns identified 204 different hospital pathways followed by the 1095 patients. However, 14 pathways were followed by more than 73% of the patients. Most of the patients were hospitalized for the same episode of AMI in at least two different units (646 patients, i.e., 59% of the study population). Those patients hospitalized in three or more units were actually transferred

from a local hospital to a tertiary level hospital such as the teaching hospital of Dijon or the private hospital healthcare institution (Clinique Fontaine). The step variable represented the chronological order of a patient's admission defined by the pair: hospital unit \times healthcare institution at a given time. These pathways fell into one of ten patterns identified in Table 2.

There are eight expert classes identified by cardiologists, and how they match patterns is shown in Table 2. In the classical analysis, these patterns/classes were considered as the modalities of a covariate, expert class. In the symbolic analysis, these patterns for each of the 8 expert classes were symbolic-valued realizations.

4.2. Regression Tree Analyses on 1095 Patients

The classical regression tree analysis on the 1095 patients produced the hierarchical tree shown in Figure 2. The most important factor is TIMI score with a cutoff value of 7.5. Regarding patients with $\text{TIMI} \geq 7.5$, expert class is the next important factor, while regarding the others, the second cut is led by TIMI score again, but with a threshold of 3.5. For the 510 patients with $3.5 \leq \text{TIMI} \leq 7.5$, the further important variables are, respectively, whether or not these patients were treated for high cholesterol, and there was a hereditary component. Expert class is the third important factor regarding patients with $\text{TIMI} < 3.5$. Most of the patients with $\text{TIMI} \geq 7.5$ (N2) have the worst survival rate (95.6%). Although TIMI is the main important factor, the pathway did, however, play a role: at the second cut for the few (84) patients for whom $\text{TIMI} \geq 7.5$ and at the third cut separating out the 25 patients who went to a noncardiology unit on admission.

4.3. Ecological Analysis on 204 Pathways

A CART analysis on the "mean" values of the 204 pathways, that is, the so-called ecological analysis, produced the tree shown in Figure 3. As for the analysis on the individual patients, TIMI score is the most important predictor variable for survival: higher TIMI scores ≥ 7.2 (N1) correspond to the poorest survival rate (52.5%) whereas lowest $\text{TIMI} < 1.75$ (N7) are associated with a perfect survival (100%). However, expert class is this time the intermediate second predictor variable for pathways having $1.75 \leq \text{TIMI} \leq 7.2$. Regarding those pathways, greatest survival rates (89.9% at N4 and 80.0% at N3) are reached for expert class 7, that is, for patients who were both admitted in cardiology and transferred to a cardiology department.

Patients in expert class 8 (pathways in which patients first went to a noncardiology unit) had a survival rate of 79.4% (terminal node N2). Further cuts then occur for expert class 7, related to heredity, with a lower survival rate (80.0%) in the case of high probability of heredity (N3) rather than in low heredity pathways (N4, 89.9%). For lower expert class pathways, the next variable is gender, with a higher proportion of males (N5) having a low survival rate at 56.7%, while women (N6) have a better but still relatively poor survival rate of 67.1%.

4.4. Symbolic Analysis on 204 Pathways

The symbolic analysis on the 204 pathways described by symbolic values, produced the classification tree shown in Figure 4. Pathways are the most important prognostic factor, with the variable expert class (equal to 7 or not) being the first discriminating factor. Expert class

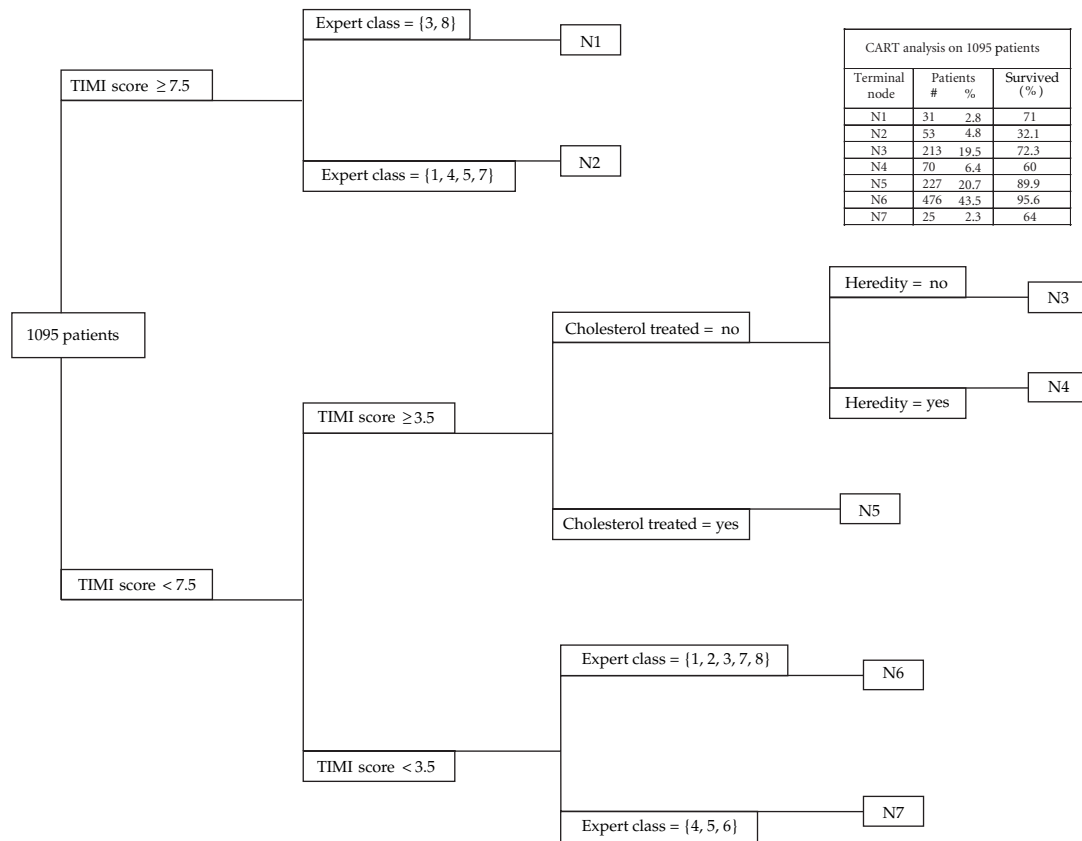


Figure 2: Discrimination tree on 1095 patients.

7 relates to pathways for which both the first and second units were cardiology units albeit it at multiple hospitals. The 77 pathways all coming from expert class 7 (N1 to N7) had a high one-year survival rate of 91.9% while the 127 pathways with a probability less than 0.69 of containing only cardiology units had an overall survival rate of 76.7%. Thus, at the outset, not only is pathway/class category important, this analysis identifies the importance of admission to cardiology units when experiencing a heart attack.

Regarding pathways with expert class 7, while the overall survival rate is 91.9%, we observe that nodes N2 and N3 have very high survival rates of 95.4% and 100%, respectively. In addition to being pathways dominated by admission to cardiology units, these also have low probability of family history of heart problems (second cut). In contrast, at node N5 the survival is only 38.9%; however, the numbers here are small (only 9 pathways and 16 patients).

While pathways with low probability of "expert class 7" have an overall survival rate of 76.7% and terminal nodes' survival rates range from 29.2% to 85.1%. The best rate here 85.1% occurred at node N13, where most pathways began in cardiology units but the numbers are relatively small.

Notice that in sharp contrast to the classical analysis on the individual patients, TIMI score does not appear as a discriminating variable until the fourth cut stage of the tree construction. Moreover, node N6 which is identified by a high probability of a low TIMI

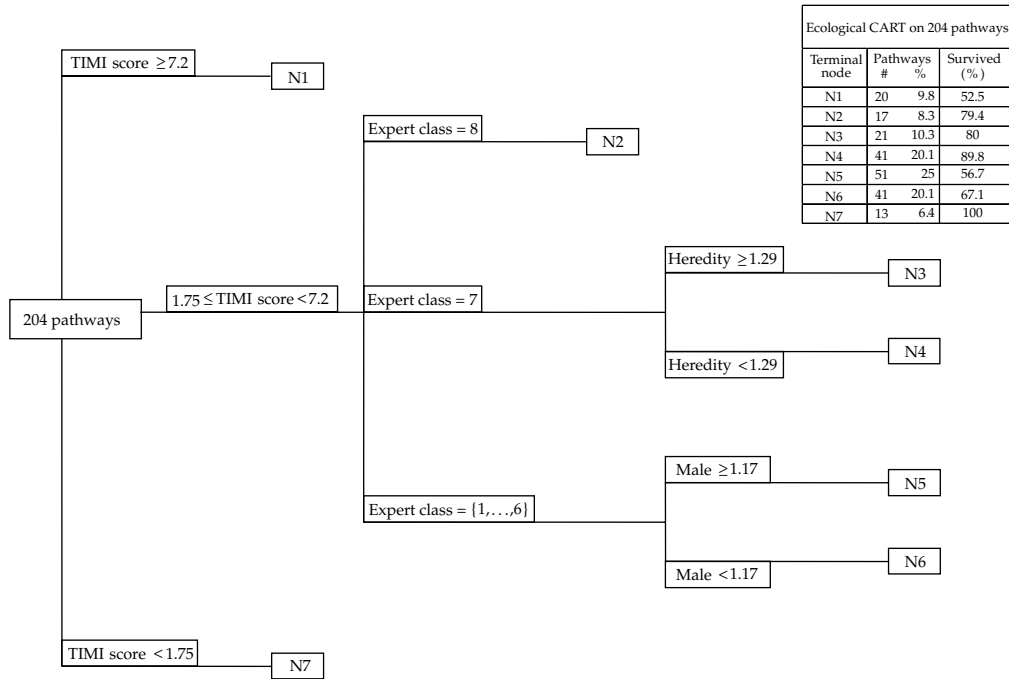


Figure 3: Ecological discrimination tree on 204 pathways.

score, has the lowest survival rate for this part of the tree; however, the numbers are very small. Actually, together the four TIMI nodes (N6, N7, N11, and N12) only accounted for 68 (6.2%) patients.

Further branches (not shown in Figure 4) gave less informative results. In this case, the discriminating variable was either TIMI score, or one of hypertension, hypercholesterol levels, ST elevation, and diabetes, which are confounded with TIMI score. The numbers are also low, exacerbating attempts at analysis and interpretation at this level.

Table 5 provides the detailed breakdown of the number of pathways in each of the expert classes at each of the terminal nodes obtained by the symbolic analysis on these 204 pathways. It also gives the corresponding numbers of patients. Thus, we see, for example, at node N2, 22 of the 28 pathways consisting of 47 of the 54 patients first went to a cardiology unit (expert class: 1, 3, and 6) and 5 pathways consisting of 6 patients (expert class 5) were transferred to a cardiology unit after first admission to a noncardiology unit. Only one pathway of one patient went to a noncardiology unit (expert class 2) before being discharged. The survival rate for this node was 95.4%.

When applying the same methodology to these 204 pathways but replacing TIMI score variable by age, the first discriminating variable is age (65-years-old threshold). For older patients, expert class (=7 or not) is the second most important predictor variable, followed by heredity and then a mix of diabetes, hypertension and ST elevation effects. For younger patients, gender is the next leading predictive variable. When both TIMI and age are included, almost the same tree structure emerges.

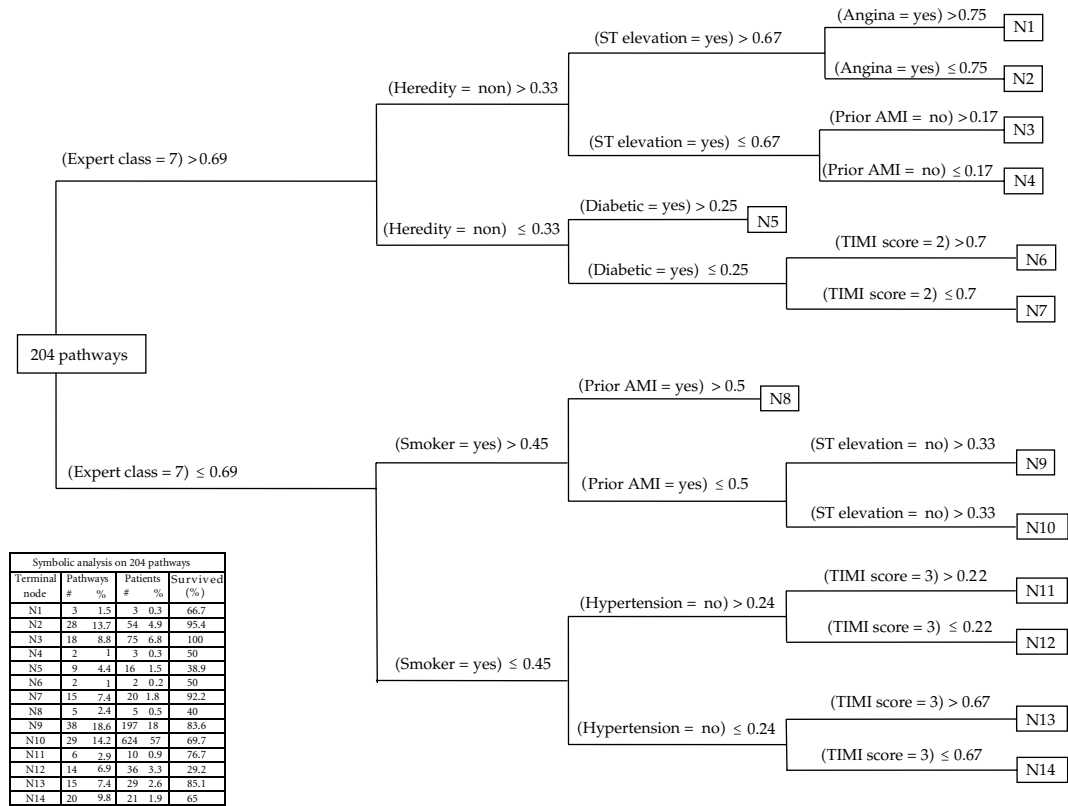


Figure 4: Symbolic discrimination tree on 204 pathways.

Table 5: Expert class-terminal node: symbolic analysis on pathways.

Node	Number of pathways									Number of patients								
	Expert class								Total	Expert class								Total
N1	.	.	1	.	.	2	.	.	3	.	.	1	.	.	2	.	.	3
N2	2	1	2	.	5	18	.	.	28	3	1	2	.	6	42	.	.	54
N3	1	2	1	.	.	14	.	.	18	9	8	20	.	.	38	.	.	75
N4	2	.	.	2	3	.	.	3
N5	.	.	1	.	.	6	.	2	9	.	.	1	.	.	13	.	2	16
N6	2	.	.	2	2	.	.	2
N7	2	1	4	.	1	7	.	.	15	2	1	4	.	1	12	.	.	20
N8	3	2	.	.	5	1	2	.	.	5
N9	.	4	4	6	3	18	.	3	38	.	19	5	15	6	149	.	3	197
N10	.	.	2	1	2	21	1	2	29	.	.	2	1	2	508	109	2	624
N11	.	1	.	.	2	3	.	.	6	.	1	.	.	6	3	.	.	10
N12	.	.	1	.	.	12	.	1	14	.	.	1	.	.	32	.	3	36
N13	.	1	2	.	.	12	.	.	15	.	1	2	.	.	26	.	.	29
N14	.	3	4	1	3	9	.	.	20	.	4	4	1	3	9	.	.	21
Σ	5	13	22	8	19	128	1	8	204	14	35	42	17	27	841	109	10	1095

5. Discussion

Comparing the results from the different analyses, we see immediately that the symbolic analysis does indeed identify pathways as an important factor. In this case, the cardiology components of pathways (since, e.g., expert class = 7 contains only cardiology units) is the first and most important discriminating prognostic variable. This feature did not emerge in the hierarchical tree when conducting the classical analysis with patients as the statistical unit (see Section 4.2) until the third cut for most (1011) of the patients, and at the second cut for those 84 patients with TIMI >7.5. These results show that the classical analysis identified the TIMI score as the most important predictor of survival, whereas the new methodology based on the pathway units identified the pathway through the expert class variable as the main predictor of the survival rate along with other relevant covariates such as smoking, heredity, diabetes, hypertension, having a prior infarction or angina, and ST-segment deviation.

A further analysis of subsequent nodes in the symbolic tree also identified higher survival rates at nodes that had predominately cardiology units at the first or second steps along the pathway. Thus, our analysis has expanded the earlier conclusions of [12–14] which looked at only the admitting physician (cardiologist or not) or of [15] which considered what in the pathway setup was the first step only, to conclusions that show the complete sequence of steps in a pathway is an important predictor of survival.

The ecological analysis, like the symbolic analysis, also considered pathways but did not use all the information contained in the pathways. The ecological analysis used only the central tendency value obtained from among all the individuals included in the respective pathways, whereas the symbolic analysis used all the information. As a consequence, the symbolic analysis is able to identify pathways as the first predictor of survival something missed in the ecological study as shown in the comparison between the ecological and symbolic analyses of the present cardiology study. The ecological analysis did not identify the pathway as being the most important (in the sense that this was not the first cutting variable selected in the tree construction), but did identify this variable at the second branching of the tree; see Section 4.3. The classical and ecological analyses started with TIMI score as their first explanatory variable. In this sense then, the ecological analysis can be viewed as having provided a more informative analysis than the classical one but still not as good as was the symbolic analysis. Thus, the symbolic analysis has all the advantages of the ecological analysis but does not have the ecological analysis' disadvantages.

From a clinical point of view, the different results in the selection of factors associated with higher in-hospital one-year mortality provided by the three classification trees are to be questioned. The classical regression tree with patients as statistical units as well as the ecological tree based on central tendency values of the pathways as the statistical unit indicated that the TIMI score was the primary discriminant variable. This result was expected in as much as the severity score was a recognized prognostic factor after AMI [37]. Surprisingly, while differences appeared in the risk of death due to disease severity at admission when patients' pathways were considered as statistical units, neither the type of healthcare institution nor the "reduced" hospital paths (as expressed through the "expert class" variable) were discriminant variables in the standard regression tree; see Figure 2. Yet, it is well known that the severity of AMI at admission will govern the process of care and especially the indication of transfer to an intensive care unit or department of cardiology at a teaching hospital [12, 15]. The ecological analysis gave some improvement with its selection of the pathway variable at the second cut of the tree. The interim nature of the ecological result is a natural consequence of the fact that these values take the central tendency values

of the symbolic values but ignore the internal variation information used in the analysis of the symbolic data.

Several explanations can be proffered to explain such discrepancies between classical and symbolic data regression trees. The main explanation may be the following. The variable “hospital pathway” can be considered as an almost intermediate variable in the relationship between AMI severity and death by AMI. Indeed, patient characteristics such as age, AMI severity or distance of the patient’s residence from a specialized care center govern the indication of hospital pathway. However, the type of hospital pathway also determines the pattern of care as well as the quickness of therapeutic interventions and, as a consequence, the outcome of acute myocardial infarction. In this case, it is difficult to consider the variable “hospital pathway” as a potential explanatory variable of death by AMI at the same level as other determinants of patient’s mortality such as age, medical history of coronary events, or comorbidities. Introducing “hospital pathway” as a “classical” independent variable along with the other independent variables in a regression model exposes the risk of over adjustment of the model. This makes it more interesting to use symbolic data analysis, since it allows for consideration of the types of hospital pathways as statistical units instead of as an explanatory variable of death by AMI. However, using types of hospital pathways as statistical units implies that independent variables in symbolic data analysis explain death by AMI within hospital pathways rather than patient death by itself.

As a consequence, these discrepancies were then somewhat expected since the interpretation of the three regression trees was different. With the classical or ecological regression trees, we determined which variables could discriminate patients according to their living status after AMI whatever their pathway. In the symbolic approach, we are interested in identifying the variables that can explain the death of patients who followed the same hospital pathway for myocardial infarction (AMI). Our symbolic analysis has then shown that pathway itself becomes an important, indeed the most important additional entity but the interpretation of the covariates’ effect on death probability is not straightforward when the statistical unit is the patient’s pathway instead of the patient. We are not used to interpreting these results, as classical analyses do not allow for the inclusion of pathways instead of patients as statistical units. A first step in this direction is proposed by the multilevel regression analysis which allows us to consider pathways. However, symbolic analysis enables us to go much further. Of course, epidemiologists may be wary at first of this new way of analysing and interpreting data.

From a methodological point of view, an advantage of the symbolic approach is that the paradigm makes it possible to create higher level statistical units described by symbolic data, therefore, enabling the scientist to settle and solve new questions. Here, the new statistical units are the pathways. A different question could be “which logical combination of biological, demographic or clinical variables discriminates pathways or subpathways (described by symbolic data) of nonsurviving patients against pathways of living patients”.

A general advantage of the symbolic approach is the reduction in the number of units. This advantage is particularly important for very large databases, where in theory, there is no problem but in practice the usual classical analyses can become problematical due to its size. Further, a result of these aggregations is that the problem of missing values is reduced considerably. Also, there is an improvement in confidentiality issues.

6. Conclusion

In this work, we have presented a new approach of data analysis in the framework of data mining: symbolic data analysis applied to medical data. Symbolic data analysis has the advantage of discovering potential associations between variables in taking advantage of the natural hierarchical structure of the data. It is achieved by the definition of new statistical units corresponding to a level of aggregation higher than the patient. It allows for the extraction of statistical information from complex data such as a complete history of hospital pathways of patients hospitalized for AMI, allowing epidemiologists to discover new relations that do not appear in, nor can they be identified by, a standard statistical analysis.

References

- [1] K. Barakat, P. Wilkinson, A. Deaner, D. Fluck, K. Ranjadayalan, and A. Timmis, "How should age affect management of acute myocardial infarction? A prospective cohort study," *Lancet*, vol. 353, no. 9157, pp. 955–959, 1999.
- [2] P. De Araújo Gonçalves, J. Ferreira, C. Aguiar, and R. Seabra-Gomes, "TIMI, PURSUIT, and GRACE risk scores: sustained prognostic value and interaction with revascularization in NSTEMI-ACS," *European Heart Journal*, vol. 26, no. 9, pp. 865–872, 2005.
- [3] A. Menotti and M. Lanti, "Coronary risk factors predicting early and late coronary deaths," *Heart*, vol. 89, no. 1, pp. 19–24, 2003.
- [4] A. Nigam, R. S. Wright, T. G. Allison et al., "Excess weight at time of presentation of myocardial infarction is associated with lower initial mortality risks but higher long-term risks including recurrent re-infarction and cardiac death," *International Journal of Cardiology*, vol. 110, no. 2, pp. 153–159, 2006.
- [5] G. D. Thakker, N. G. Frangogiannis, M. Bujak et al., "Effects of diet-induced obesity on inflammation and remodeling after myocardial infarction," *American Journal of Physiology*, vol. 291, no. 5, pp. H2504–H2514, 2006.
- [6] G. Casella, S. Savonitto, F. Chiarella et al., "Clinical characteristics and outcome of diabetic patients with acute myocardial infarction. Data from the BLITZ-1 study," *Italian Heart Journal*, vol. 6, no. 5, pp. 374–383, 2005.
- [7] M. Kamalesh, U. Subramanian, A. Ariana, S. Sawada, and W. Tierney, "Similar decline in post-myocardial infarction mortality among subjects with and without diabetes," *American Journal of the Medical Sciences*, vol. 329, no. 5, pp. 228–233, 2005.
- [8] G. Hu, P. Jousilahti, Q. Qiao, S. Katoh, and J. Tuomilehto, "Sex differences in cardiovascular and total mortality among diabetic and non-diabetic individuals with or without history of myocardial infarction," *Diabetologia*, vol. 48, no. 5, pp. 856–861, 2005.
- [9] M. G. Abrignani, L. J. Dominguez, G. Biondo et al., "In-hospital complications of acute myocardial infarction in hypertensive subjects," *American Journal of Hypertension*, vol. 18, no. 2, part 1, pp. 165–170, 2005.
- [10] K. A. Eagle, M. J. Lim, O. H. Dabbous et al., "A validated prediction model for all forms of acute coronary syndrome estimating the risk of 6-month postdischarge death in an international registry," *Journal of the American Medical Association*, vol. 291, no. 22, pp. 2727–2733, 2004.
- [11] J. C. Beer, G. Dentan, L. Janin-Magnificat et al., "Beneficial effects of direct call to Emergency Medical Services on time delays and management of patients with acute myocardial infarction. The RICO (observatoire des Infarctus de la Côte-d'Or) data," *Annales de Cardiologie et d'Angiologie*, vol. 51, no. 1, pp. 8–14, 2002.
- [12] P. N. Casale, J. L. Jones, F. E. Wolf, Y. Pei, and L. M. Eby, "Patients treated by cardiologists have a lower in-hospital mortality for acute myocardial infarction," *Journal of the American College of Cardiology*, vol. 32, no. 4, pp. 885–889, 1998.
- [13] J. G. Jollis, E. R. Delong, E. D. Peterson et al., "Outcome of acute myocardial infarction according to the specialty of the admitting physician," *New England Journal of Medicine*, vol. 335, no. 25, pp. 1880–1887, 1996.
- [14] M. F. Dorsch, R. A. Lawrance, R. J. Sapsford et al., "An evaluation of the relationship between

- specialist training in cardiology and implementation of evidence-based care of patients following acute myocardial infarction," *International Journal of Cardiology*, vol. 96, no. 3, pp. 335–340, 2004.
- [15] M. Gottwik, R. Zahn, R. Schiele et al., "Differences in the treatment and outcome of patients with acute myocardial infarction admitted to hospitals with compared to without departments of cardiology. Results from the pooled data of the Maximal Individual Therapy in Acute Myocardial Infarction (MITRA1+2) Registries and the Myocardial Infarction Registry MIR," *European Heart Journal*, vol. 22, pp. 1794–1801, 2001.
- [16] M. H. Gail, S. Wieand, and S. Piantadosi, "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates," *Biometrika*, vol. 71, no. 3, pp. 431–444, 1984.
- [17] N. Keiding, P. K. Andersen, and J. P. Klein, "The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates," *Statistics in Medicine*, vol. 16, no. 1–3, pp. 215–224, 1997.
- [18] C. Schmoor and M. Schumacher, "Effects of covariate omission and categorization when analysing randomized trials with the Cox model," *Statistics in Medicine*, vol. 16, no. 1–3, pp. 225–237, 1997.
- [19] R. A. Betensky, D. N. Louis, and J. G. Cairncross, "Influence of unrecognized molecular heterogeneity on randomized clinical trials," *Journal of Clinical Oncology*, vol. 20, no. 10, pp. 2495–2499, 2002.
- [20] L. Billard, "Sample covariance functions for complex quantitative data," in *Proceedings World Congress International Association Statistical Computing*, 2008.
- [21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.
- [22] O. Montagne, C. Chaix, A. Harf, A. Castaigne, and I. Durand-Zaleski, "Costs for acute myocardial infarction in a tertiary care centre and nationwide in France," *Pharmacoeconomics*, vol. 17, no. 6, pp. 603–609, 2000.
- [23] D. A. Morrow, E. M. Antman, A. Charlesworth et al., "TIMI risk score for ST-elevation myocardial infarction: a convenient, bedside, clinical score for risk assessment at presentation: an Intravenous nPA for Treatment of Infarcting Myocardium Early II trial substudy," *Circulation*, vol. 102, no. 17, pp. 2031–2037, 2000.
- [24] G. Davey Smith and A. N. Phillips, "Confounding in epidemiological studies: why 'independent' effects may not be all they seem," *British Medical Journal*, vol. 305, no. 6856, pp. 757–759, 1992.
- [25] M. Egger, M. Schneider, and G. D. Smith, "Meta-analysis spurious precision? Meta-analysis of observational studies," *British Medical Journal*, vol. 316, no. 7125, pp. 140–144, 1998.
- [26] J. M. Oakes, "The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology," *Social Science and Medicine*, vol. 58, no. 10, pp. 1929–1952, 2004.
- [27] J. S. Kaufman, R. S. Cooper, and D. L. McGee, "Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race," *Epidemiology*, vol. 8, no. 6, pp. 621–628, 1997.
- [28] H.-H. Bock and E. Diday, Eds., *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Germany, 2000.
- [29] L. Billard and E. Diday, "From the statistics of data to the statistics of knowledge: symbolic data analysis," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 470–487, 2003.
- [30] L. Billard and E. Diday, *Symbolic Data Analyses: Conceptual Statistics and Data Mining*, Wiley Series in Computational Statistics, John Wiley & Sons, Chichester, UK, 2006.
- [31] G. C. Fonarow, K. F. Adams, W. T. Abraham, C. W. Yancy, and W. J. Boscardin, "Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis," *Journal of the American Medical Association*, vol. 293, no. 5, pp. 572–580, 2005.
- [32] K. R. Hess, M. C. Abbruzzese, R. Lenzi, M. N. Raber, and J. L. Abbruzzese, "Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma," *Clinical Cancer Research*, vol. 5, no. 11, pp. 3403–3410, 1999.
- [33] L. Billard and E. Diday, "Regression analysis for interval-valued data," in *Data Analysis, Classification, and Related Methods*, H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader, Eds., Studies in Classification, Data Analysis, and Knowledge Organization, pp. 369–374, Springer, Berlin, Germany, 2000.
- [34] L. Billard and E. Diday, "Symbolic regression analysis," in *Classification, Clustering, and Data Analysis*, K. Jajuga, A. Sokolowski, and H.-H. Bock, Eds., pp. 281–288, Springer, Berlin, Germany, 2002.
- [35] F. Afonso, L. Billard, E. Diday, and M. Liman, "Symbolic linear regression methodology," in *Symbolic Data Analysis and the SODAS Software*, E. Diday and M. Noirhomme-Fraiture, Eds., pp. 359–372, John

Wiley & Sons, Chichester, UK, 2008.

- [36] D. Seck, L. Billard, E. Diday, and F. Afonso, "A decision tree for interval-valued data with modal dependent variable," *COMPSTAT*, vol. 19, pp. 1621–1628, 2010.
- [37] M. Singh, G. S. Reeder, S. J. Jacobsen, S. Weston, J. Killian, and V. L. Roger, "Scores for post-myocardial infarction risk stratification in the community," *Circulation*, vol. 106, no. 18, pp. 2309–2314, 2002.