# Trend analysis and risk identification

**Novakova L.*, Klema J.*, Jakob M.*, Rawles S.**, Stepankova O.***

*Gerstner Laboratory, Department of Cybernetics, FEE, Czech Technical Univerzity, Prague, Czech Republic.

**Department of Computer Science, Bristol University, Bristol, UK.

**Abstract.** The 2003 ECML/PKDD data mining challenge concerns a dataset describing the data collected during a longitudinal study of atherosclerosis prevention on around 1400 middle-aged men at a number of Czech hospitals. The data challenge entry from the Czech Technical University in Prague takes an approach which is heavy on data preparation through well-defined data transformations. This document describes the special requirements of this data mining tasks, the transformations designed to meet them and it points to some interesting observations found in the studied dataset.

## 1 The data, tasks and tools

### 1.1 Overview

Subject of this contribution is the STULONG data set concerning the twenty years lasting longitudinal study of the risk factors of the atherosclerosis in the population of 1 417 middle aged men. The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomečková, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvárová, DrSc). At present time the data analysis is supported by the grant of the Ministry of Education CR Nr. LN 00B 107.

The data is inherently multi-relational, consisting of four separate tables. The table Entry describes data collected during the entry examinations of all patients, Control includes results of a series of long-term observations recording the development of risk factors and associated conditions, Letter provides complementary information collected by questionnaire filled-in by all the patients and records about death of some patients appear in the Death table.

### 1.2 Data exploration

The first step of the domain analysis was a data exploration phase. The cleanliness of the data was verified, and basic understanding of the data was achieved

using appropriate modules in SumatraTT 2.0. The first-order Bayesian classifier Tertius [4] was used in an exploratory way to determine which of the attributes have dependancies on which other attributes. The results have been compared to the findings concerning a similar dataset, which have been summarized in the proceedings of the Data Challenge 2002.

### 1.3 The tasks and tools

Most of the participants of the Data Challenge 2002 treated the four data tables separately. For this reason we have decided to focus on the task which would allow us to take into account all the data supplied, that is, both the entry examinations and the long-term observation. Our intention is to provide an answer to the question: *Is there any difference in the development of risk factors and other characteristics between men of the risk group who came down with the observed cardiovascular diseases and those who stayed healthy?*

We have planed to approach the task as a learning problem by transforming the multi-relational domain first to a propositional case, and then to induce models using an attribute-value learner. There are two reasons for taking this approach. First of all it allowed us to test the available features of the data processing tool SumatraTT 2.0 [3, 1], which is well suited to producing rather complex merges of various tables. Moreover, there exists a large number of proven learners on propositional datasets. Our learning experiments have relied on Weka and Statistica [2].

### 1.4 Problems appearing in the data

Our approach places a lot of emphasis on data transformation [7]. The process of analysing the data led to a series of special requirements and challenges associated with the data in its supplied form. These are summarised below:

*Multi-relational with 1-to-n cardinalities.* The data consists of four tables and is inherently multi-relational. Furthermore, there are records for each examination in the longitudinal study relating to each patient, introducing a cardinality of 1-to-$n$ between these tables. This prompts the use of carefully-considered and domain-specific aggregation techniques in order to represent these multiple records in a single table with minimal information loss.

*Time series data.* The exam histories for each patient form a sparse time series data set. When aggregating, suitable techniques for time series analysis need to be chosen in order to best summarise the data. [6]

*Patients have varying medical histories.* The exam histories of each patient can differ. For example, the start and end dates of the periods of study can differ, as well as the length of time the patient was studied for and the frequency and regularity of examinations.

*Varying patients.* The patients themselves are of different ages at comparable points in the medical history. Differences in age and factors like age make different patients less comparable.

*Cause-and-effect prediction task.* Predicting the onset of diseases and symptoms from the factors identified in the study requires a treatment of the data beyond taking the full history and predicting from that. For the induced rules to be useful, they need to be expressed in such a way that they allow prediction of the onset of disease *given information about a current situation only.* In other words, for the induced models to make sense, they must predict the effect given only the causes.

*Sparse data.* The data is relatively clean, but it is of course impractical for the clinicians to perform every possible test on every patient in each exam. Therefore the data has missing values where tests or measurements were not completed, and this makes some attributes very sparse.

## 2 The transformations and derived attributes

Firstly, we transform the data within each table to suit its role within the task and tools to be used. This involves representing data in a useful way to the learner, but also aiming for appropriate transformations, so that there is both maximal exposure of information content [6] and the output rules make sense. These transformations include simple operations such as attribute selection and removal, to more complex operations such as clustering and discretization.

An important feature of the Contr table is that each of the patients appears several times — the number of examinations of a single patient ranges from 1 to 20. There are even (about 60) patients in Entry table, who do not appear in the Contr table at all. These patients were removed and no longer considered. Specific types of aggregation have been designed to create a single record on each of the patients - see section 2.2, Handling the exam history of a patient.

Finally, straightforward multi-relational data transformations offered by SumatraTT 2.0 [3] have been applied to produce a single table merging data from the enhanced Entry and Contr tables.

### 2.1 Initial transformations

First, the data in each table have to be transformed to suit the considered task or to contribute to the understandability of the models induced.

*Attribute selection and removal.* This transformation simply removes attributes which are not of interest to the data mining task, either because they are too sparse or because they are irrelevant, for example attribute CHLSTMG (cholesterol in mg%) and CHLST (cholesterol in mmol/l), the difference between these two attributes only existing in the unit used.

*Discretization.* A number of the attributes appeared as categorical or nominal values. In order to make them comparable with other attributes we used a transformation to re-categorise these values to make their values the same as other, comparable attributes.

*Decimal time.* In order to compare and normalise times in the data, the decimal time transformation converts times specified as years, months to a single floating point number, for example $newtimeattr = year + \frac{1}{12}\ month$.

*Derived attributes.* We designed two new attributes - BMI and Disease.

$$BMI = \frac{weight[kg]}{height^2[m]} \tag{1}$$

All patients in the Contr table who suffered from HODN4 or HODN12 or HODN15 and having no cardiovascular disease, were removed from the table Entry. The resulting table was named Entry1.

We introduce CVD (meaning cardiovascular disease), a new boolean attribute for patients in the table Entry1. If it is false the patient has no coronary disease. If it is true it means that the patient has some cardiovascular disease, namely at least one of the entries in the Contr table for the considered patient has the value true for one of the following attributes: HODN1, HODN2, HODN3, HODN11, HODN13, HODN14, HODN21, HODN23.

### 2.2 Handling the exam history of a patient

In order to produce a single table from the long-term observation data for each patient in the Contr table, we implemented a module to calculate derived attributes describing the trends in the measurements taken (e.g. BMI, systolic and diastolic blood pressure, cholesterol, triglyceride level, smoking). There were two options for how to proceed. The first possibility is to apply regression to data from all available exams – the global approach. An alternative way is to restrict the data to a limited time window. Both these approaches (described in more detail below) apply following aggregations.

*Use of aggregates.* Some of the attributes from the long-term observation table allow aggregation by using simple statistical measures such as the mean, standard deviation and counts of particular values. This is handled by the statistical module in SumatraTT 2.0.

*Area under a graph.* Some of the factors make sense when considered over a period of time. An example of this is smoking of cigarettes. Consider a graph of the number of cigarettes smoked per day over time. By calculating the area under this graph we can get an estimate of the total number of cigarettes smoked in the period bounding that area. This estimate can be used as an additional derived attribute to give both a long-term and short-term measure of the amount smoked.

**Global approach** The following five derived attributes were calculated to produce trend attributes for all the continuous original attributes.

- The mean and standard deviation of the non-time variable.
- For a linear fit of the data, the correlation coefficient $r$ and the gradient $m$ and $y$-intercept $b$ of the best-fit line.

The best fit [5] is calculated by pre-calculating the statistical values of $n$, $\sum x$, $\sum y$, $\sum x^2$, $\sum xy$, $\sum y^2$.

The y-intercept is calculated as $b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$.

The gradient $m$ is calculated as $m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$.

The correleation coefficient $r$ is calculated as $r = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$.

**Windowing** Differences in the number of exams taken by different patients can cause serious side effects. To eliminate this danger we have decided to restrict the attention to a comparable part of the data provided for each patient, namely, only data from a time windows of the fixed number of consecutive measurements have been taken into account for each patient. The length of this time window has to be the same for all the patients - number five has been chosen because most of the patients have at least five entries in the Contr table - see Figure 1. The patients with less than five entries in the Contr table or those for whom most measurements were lacking have been removed from the further study. The resulting table is referred to as Contr1.

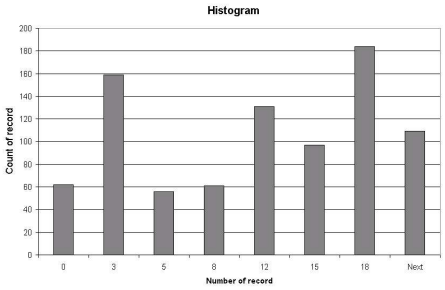



**Fig. 1.** Histogram of measurement count for each patient in Contr table

The aggregates based on the constant number of measurements deal with a window of fixed length corresponding to five measurements sliding from the first to the last measurement for each patient, i.e., one patient is represented by more than one row. As the time between the first and the fifth measurement can differ, it is stored as a normalizing parameter. The dependent variable is the time between the end of the current window position and the possible onset of disease. In the case that the patient stays healthy, the value 0 is stored. A patient with 20 measurements taken once a year whose disease was discovered at the last measurement will be represented by following 15 records:

m1_value, ..., m5_value, 5 years, 15 years
m2_value, ..., m6_value, 5 years, 14 years
...
m15_value, ..., m19_value, 5 years, 1 year

The trend attributes have been calculated for the time window in the same way as in the global case. The mean value of all five measurements was also computed for each time window.

The windowed aggregates were constructed for the following attributes: cholesterol, BMI, smoking, triglycerides, systolic and diastolic blood pressure. The linear trends were normalized by the period between the first and the fifth measurement.

## 3 Modelling based on the Entry table

Prior to the analysis of the derived attributes, we analyzed the Entry table, in order to gain better understanding of individual attributes and their relation to the occurrence of a CVD. Thus, as a subgoal, we first tried to answer analytical question number 6: *Are there any differences in the entry examination between men of the risk group, who came down with the observed cardiovascular diseases and those who stayed healthy?*

Application of several machine learning techniques revealed that the attributes in the Entry table have relatively weak predictive power. The highest predictive accuracy as measured by 10-fold cross-validation experiments was in fact achieved by trivial classifiers classifying all examples in class zero, i.e. as healthy. In order to improve the classification of class one cases (i.e. patients which came down with a CVD), we further experimented with cost-sensitive learners with an asymmetric error-cost matrix. Although we applied a variety of different learners — including decision tree and decision rule induction algorithms, multilayer perceptrons and Bayes classifiers – none of which achieved satisfactory results.

Having come to this piece of knowledge, we abandoned the original idea of building a model, which would classify each patient in one of the two possible CVD classes. Instead, we focused on the identification of factors related to the increased rate of a CVD. For each of such risk factors, we tried to identify subgroups of patients showing significantly different rates of CVD in comparison to the overall CVD rate.

### 3.1 Discovery of Interesting Subgroups

We applied the Statistica 6.0 [2] module for interactive decision tree induction to fulfill the above given goal. We were interested mainly in groups describable by a small number of attributes. In such cases, the interactivity and flexibility of the system made this approach a viable alternative to a specialized subgroup discovery system.

Each of the identified subgroups is defined by a condition on one or two attributes, see Table 1-4. Names and values of attributes values corresponding to individual subgroups are presented in the first columns of subgroup description. The next two columns display the rate of CVD in the subgroup and its size, respectively. For each subgroup, we have performed a two-tailed $t$-test to assess

whether the CVD rate within the group is statistically significantly different to the overall CVD rate. The level of significance $p$ is depicted in the last column of subgroup description.

### 3.2 Subgroups Based on the Entry Table

The input to the analysis was the preprocessed Entry, which includes only those patients from the original Entry table who are from the risk group, and who stayed healthy or came down with a cardiovascular disease (see Figure 2).

We do not cover a well-known risk factors here – such as smoking, BMI and cholesterol level – and instead focus on less known, perhaps even surprising dependencies we have discovered. As already mentioned, our goal was to identify subgroups with CVD rate significantly different to the base CVD rate.

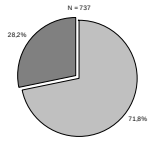



**Fig. 2.** The Entry table after preprocessing. The light segment represents patients from the risk group who stayed healthy, the dark one represents patients from the risk group who suffered a CVD.

*Social Characteristics.* We have discovered in the Entry table that age is the strongest factor correlated with CVD (correlation with CVD is $r = 0.15$). Because it is not directly represented in the Entry table, we included it in this survey even though it is certainly a well-known factor. In order to analyze the influence of age on CVD occurrence, we introduced a new attribute VEK := ROKVSTUP – ROKNAR. The relation of the age with CVD rate is clearly demonstrated in Figure 3 and Table 1.

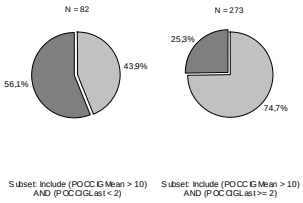



**Fig. 3.** Dependance of the CVD rate on the age of patient at the time of entry into the study.

Vertical bars shows 95% confidence intervals for estimates of CVD rate.

*Alcohol.* The most significant attribute in the Alcohol group is PIVOMN (correlation with CVD $r = -0.11$). The relationship between beer consumption

| VEK | CVD % | N= | p= |
|---|---|---|---|
| $\leq 42$ | 18.4 | 152 | 0.0129 |
| $(42, 48]$ | 26.1 | 360 | 0.5077 |
| $> 48$ | 38.5 | 221 | 0.0030 |

**Table 1.** Subgroups w.r.t. the age of patients

and the CVD rate is indicated in Table 2. What is more, the positive effect of beer drinking is in line with the positive effect (w.r.t. to cardiovascular diseases) of alcohol in general (see Table 3). Regular drinkers consuming over 1 liter of beer a day show the lowest rate of CVD! Contrary to wide-spread belief, we have not discovered any influence of wine consumption on the CVD rate, though.

| PIVOMN* | CVD % | N= | p= |
|---|---|---|---|
| 0 | 33.5 | 215 | 0.113 |
| 1 | 27.9 | 423 | 0.96 |
| 2 | 16.9 | 89 | 0.0235 |

*0 - does not drink beer, 1 - up to l liter a day, 2 - more than 1 liter a day

**Table 2.** Subgroups w.r.t. the consumption of beer.

| PIVOMN | ALKOHOL* | CVD % | N= | p= |
|---|---|---|---|---|
| 0 | 0 | 40.9 | 66 | 0.273 |
| 1 | 1 | 23.4 | 252 | 0.155 |
| 2 | 2 | 15.2 | 66 | 0.025 |

*0 - does not drink alcohol, 1 - drinks occasionally, and 2 - drinks regularly.

**Table 3.** Subgroups w.r.t. to the consumption of beer and alcohol in general.

*Sugar, coffee and tea.* We have found that the consumption of sugar is related to the CVD rate (Table 4a)

*Others.* We have discovered strong relationship between the CVD rate and the HYPLIP attribute from the personal anamnesis group of attributes (see Table 4b).

## 4 Modelling using the derived attributes

### 4.1 Modelling using the global approach

The utility of all the aggregate and time series transformations was evaluated. The first test was based on the chi-square test of goodness of fit. The test evaluated whether the patients divided into several categories according to a value of the derived attribute (statistical aggregates, trends, areas under graph). This test shows a different than uniform rate of risk of coming down with the observed cardiovascular diseases. The test has proven that a certain range of derived attributes influences the risk mentioned above at the 0.05 or 0.1 level of significance. The trend attributes (namely the linear trends of cholesterol, triglycerides and BMI) seemed to be the strongest risk factors.

| CUKR* | CVD % | N= | p= |
|---|---|---|---|
| $\leq 10$ | 27.0 | 662 | 0.676 |
| $> 10$ | 47.7 | 44 | 0.0053 |

*the number of sugar lumps a day

a)

| HYPLIP | CVD % | N= | p= |
|---|---|---|---|
| positive | 50.0 | 22 | 0.0248 |
| negative | 27.4 | 419 | 0.771 |

b)

**Table 4.** Subgroups w.r.t. a) the consumption of sugar, b) the presence of HYPLIP

However, one more group of risk factors was identified. By far the strongest risk factor proved to be the low number of control examinations (ControlCount) of a patient. The correlation attribute between ControlCount and CVD is $-0.21$. This strong relationships is subsequently transferred to all attributes closely related to the number of examinations. These numbers are not equal for all measurements (e.g., weight can be measured at different control examinations than cholesterol) but they correlate with the overall number of controls (ControlCount). Their influence on the risk of cardiovascular disease was proven at the 0.005 level of significance. The more controls the patient went through the smaller is the risk of disease - see Figure 4. This observation is not surprising, as the usual study scenario is that exactly the last control examination discovers a cardiovascular disease. Consequently, the patients who never came down with any cardiovascular disease are more likely to have a higher number of control examinations.

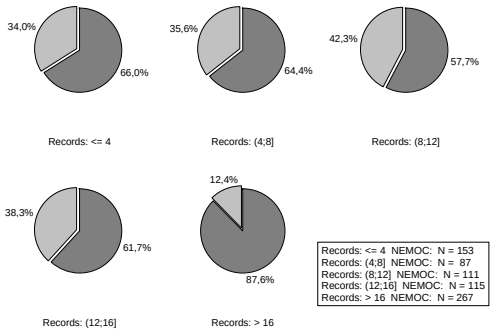
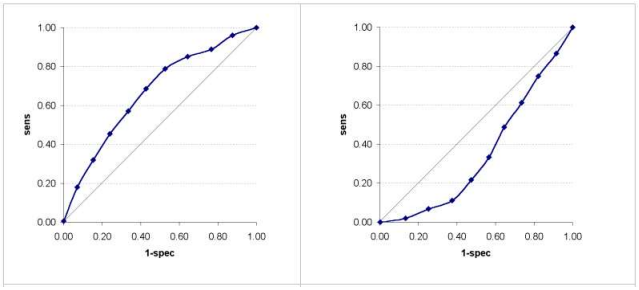



**Fig. 4.** Relation between the number of examination records and the Disease attribute

There is no doubt that the attributes directly based on number of controls (ControlCount and its derivatives) represent anachronistic attributes that cannot be practically utilized as they do not represent the cause-and-effect causality. What is more, it has been proven that a great majority of the trend attributes depends strongly on the number of controls. It can be seen that the patients with a low number of controls tend to show either a very steep trend or a zero trend. The patients with a high number of controls tend to have a mild trend. Although the trend and other derived attributes based on different number of controls show the significant influence on the final cardiovascular disease risk,

it is likely that it is only a hidden influence of ControlCount. A Naive Bayes predictor based on the number of triglycerides measurements only shows a ROC area of 0.64. If the gradient of triglycerides is added, the area grows to 0.66 only - the gradient brings little additional information).

In order to understand the above mentioned influence, it can be compared with other three well-known risk factors - Age, BMI and Cholrisk. The ROC analysis gives the following areas under the curve: Age - 0.59, BMI - 0.58, Cholesterol - 0.59, ControlCount - 0.35. It follows that ControlCount represents a stronger risk factor than Age, BMI or Cholrisk. Its predictive power is approximately as high as the power of Naive Bayes predictor combining all the three abovementioned risk factors (see Figure 5.).




a) b)

**Fig. 5.** a) ROC, Naive Bayes predictor, inputs (Age, BMI, Cholrisk), ROC area=0.67, b) ROC, ControlCount (number of examinations), ROC area=0.35

Note that the problem of anachronism plagues also aggregates officially offered by the Discovery Challenge web site under the Data Transformation section. This is because, similarly to our trend attributes, there is a strong relationship between ControlCount and the magnitude of trend attributes (see Table 5 for some of the correlation coefficients). And, as already mentioned, the design of the study introduces strong relationship between the ControlCount and CVD. Much of the predictive power of trend attributes is therefore due to the design of the study rather than due to actual casual relationships between trend attributes and CVD.

| Trend Attribute | $r$ |
|---|---|
| Syst_trend_b1_ABS | -0.34 |
| Chlst_trend_b1_ABS | -0.34 |
| POCCIG_Trend_b1_ABS | -0.28 |
| POCCIG_TrendType_ABS | -0.15 |

**Table 5.** Correlation between selected trend attributes and the number of control examinations.

The conclusion is that regarding any aggregates based on several entries in the Contr table, the main focus should be put on the aggregates based on a

constant number of measurements. This task is considered in the windowing approach.

### 4.2 Modelling based on windowing

The obtained results support the causality doubts mentioned in the previous chapter. The influence of the number of controls was minimized and consequently the relationship between the trend values and CVD occurrence has significantly weakened. For subsequent analysis, we have reformulated our question in the following way: *based on the values in the time window, predict whether the patient will suffer a CVD in three years following the last examination in the window.*

For each time window, we have thus created a new binary attribute CVD-in-3yrs. CVD-in-3yrs has the value 1 if the given patient came down with a cardiovascular disease within three years after the last examination and 0 otherwise. As an example, consider a time window consisting of five measurements in years 1978, 1979, 1980, 1981, and 1982. Then the value CVD-in-3yrs is 1 if the patient to which the time window belongs came down with the cardiovascular disease in years 1982–1985 (but not 1986 or later).

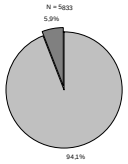



**Fig. 6.** Windowing datasets. The dark segment represents time windows where the patient has suffered a CVD three years from the last observation, i.e. CVD_in_3yrs = '1'

After such reformulation, the structure of the windowing dataset is depicted by Figure 6. Four out of a total ten windowed aggregates have been identified as related to CVD-in-3yrs: POCCIG-Mean, POCCIG-Grad, SYST-MEAN and CHLSTMG-Mean. Subgroups related to these attributes are described in Tables 6, 7, 8 and 9.

| POCCIG-Mean | CVD-in-3yrs % | N= | p= |
|---|---|---|---|
| $= 0$ | 4.7 | 2791 | 0.024 |
| $(0, 5]$ | 5.6 | 515 | 0.781 |
| $(5, 10]$ | 6.0 | 470 | 0.930 |
| $(10, 15]$ | 8.4 | 583 | 0.0120 |
| $(15, 20]$ | 8.2 | 894 | 0.008 |
| $> 20$ | 5.9 | 580 | 0.923 |

**Table 6.** Subgroups w.r.t. average number of cigarettes during the last 5 examinations.

| POCCIG-Grad | CVD-in-3yrs % | N= | p= |
|---|---|---|---|
| $\leq -5$ | 10.4 | 201 | 0.0040 |
| $(-5, 0]$ | 7.2 | 2073 | 0.0356 |
| $> 0$ | 5.5 | 768 | 0.6571 |

**Table 7.** Subgroups w.r.t gradient of cigarettes over the last 5 examinations. Only records with positive POCCIG-Mean are included.

| CHLSTMG-Mean | CVD-in-3yrs% | N= | p= |
|---|---|---|---|
| $\leq 190$ | 3.0 | 533 | 0.0056 |
| $(190, 260]$ | 5.4 | 3622 | 0.308 |
| $> 260$ | 9.5 | 1025 | $< 0.0001$ |

**Table 8.** Subgroups w.r.t. average cholesterol level during the last 5 examinations.

| SYST-Mean | CVD-in-3yrs % | N= | p= |
|---|---|---|---|
| $\leq 115$ | 2.9 | 552 | 0.0035 |
| $(115, 150]$ | 5,8 | 4613 | 0.823 |
| $(150, 160]$ | 7.8 | 475 | 0.0947 |
| $> 160$ | 12.7 | 189 | 0.0001 |

**Table 9.** Subgroups w.r.t. average number of systolic blood pressure during the last 5 examinations.

### 4.3 Subgroups Based on Windowed Aggregates

The results presented in Tables 6, 8 and 9 agree with domain knowledge concerning CVD risk factors. On the other hand, the relationship depicted in Table 7 is rather surprising. It says that the rate of CVD increases if a patient tends to stop smoking! Actually, a similar result was obtained also from the analysis based on the global approach. Namely, we have identified a subgroup (described in Table 7), which seems to claim "a patient who has given up smoking is at much greater risk than those who have not". This is, however, in contradiction with the domain knowledge. Perhaps a more plausible explanation is that patients stop smoking because their health condition becomes bad, but it is already to late to stop the coming disease.

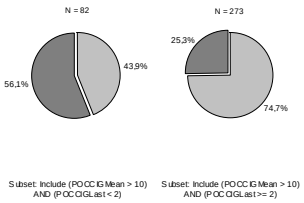



**Fig. 7.** Influence of giving up smoking on CVD occurrence.

## 5 Conclusions

We have argued in section 4.1 why the derived trend attributes based on global approach should not be used for answering the question specified in section 1. We have succeeded to identify some surprising subgroups (Table 2,3) and Table 8. In the last section we have analysed the impact of the trend attributes based on the windowing approach. It seems that none of these attributes on its own is able to identify any knowledge which could play a role in preventing CVD. The trend attributes have to be combined for the next step of analysis. Unfortunately, our choice of window (5 last controls) proved to be a wrong one as the control examinations do not have uniform structure (some tests are skipped sometimes). The correct preprocessing step has to be based on a window of fixed time period.

**Acknowledgements** The presented work was supported by the following grants: EC project ICA1-1999-75029 MIRACLE, grant of the Czech Ministry of Education MSM 210000012, grant of CTU 10-83063/2003.

## References

1. CTU, SumatraTT Homepage, 2003. krizik.felk.cvut.cz/Sumatra.
2. StatSoft,Inc. STATISTICA DataMiner Homepage, 2003. www.statsoft.com
3. P. Aubrecht, F. Železný, P. Mikšovský, and O. Štěpánková. SumatraTT: Towards a universal data preprocessor. In *Cybernetics and Systems 2002*, volume II, pages 818–823, Vienna, 2002. Austrian Society for Cybernetics Studies.
4. P. Flach and N. Lachiche. 1BC: A first-order Bayesian classifier. In S. Džeroski and P. Flach, editors, *ILP99*, volume 1634 of *LNAI*, pages 92–103. SV, 1999.
5. J. F. Kenney and E. S. Keeping. *Mathematics of Statistics*, volume 1. Princeton, third edition, 1962.
6. D. Pyle. *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, 1999.
7. O. Štěpánková, P. Aubrecht, Z. Kouba, P. Mikšovský *Preprocessing for Data Mining and Decision Support* in *Data Mining and Decision Support: Integration and Collaboration.* edited by D. Mladenič, N. Lavrač, M. Bohanec and S. Moyle. Kluwer, to appear 2003.